uc3m | Universidad Carlos III de Madrid

TESIS DOCTORAL

Contribución al Análisis de Datos de Sensores en el Ámbito de Ciudad Inteligente

Autora:

Ramona Ruiz Blázquez

Director/es:

Mario Muñoz Organero

Catedrático de Universidad

Director del Departamento de Ingeniería Telemática

Luis Sánchez Fernández

Catedrático de Universidad

Tutor:

Mario Muñoz Organero

PROGRAMA DE DOCTORADO EN INGENIERÍA TELEMÁTICA

Leganés, Abril, 2.018

Universidad
Carlos III de Madrid
www.uc3m.es

Tesis Doctoral
Ingeniería Telemática

Contribución al Análisis de Datos de Sensores en el Ámbito de Ciudad Inteligente

Autora

Ramona Ruiz Blázquez

Directores

Mario Muñoz Organero
Luis Sánchez Fernández

Departamento de Ingeniería Telemática
Escuela Politécnica Superior

2018

uc3m | Universidad Carlos III de Madrid

Tesis doctoral: **Contribución al Análisis de Datos de Sensores en el Ámbito de Ciudad Inteligente**

Autora: Ramona Ruiz Blázquez

Directores: Mario Muñoz Organero

Luis Sánchez Fernández

El tribunal nombrado para juzgar el Proyecto arriba indicado, compuesto por los siguientes miembros:

Presidente:

Vocales:

Secretario:

Acuerdan otorgarle la calificación de:

Leganés, 2018

El Secretario del Tribunal

A mi madre,

Que en paz descanse.

Agradecimientos

Quisiera expresar mi más sincero agradecimiento a los profesores Mario Muñoz Organero y Luis Sánchez Fernández por la confianza que han depositado en mí y la oportunidad que me han brindado para poder llevar a cabo esta tesis.

En segundo lugar me gustaría agradecer a la profesora Natividad Martínez Madrid de la Universidad de Reutlingen, por acogerme en su grupo, al igual que a Thomas Walzer por ofrecerme la ayuda necesaria para la consecución de mi trabajo.

También quisiera dar las gracias a Salvador Benavent Martínez por su valiosa colaboración en la recogida de datos, sin los cuales este trabajo no podría, en parte, haberse realizado, así como a todos mis compañeros y amigos por sus inestimables ayudas y consejos, y a todas aquellas personas que de alguna u otra manera han hecho de este proyecto una realidad.

Y por último, no quiero olvidar agradecer a toda mi familia por el apoyo que siempre me han mostrado.

Este trabajo ha sido financiado por el Ministerio de Economía, Industria y Competitividad, con la ayuda FPI: BES-2014-070462

Ramona Ruiz Blázquez

Leganés, 2018

Resumen

Este trabajo se enmarca dentro del vasto contexto de Ciudades Inteligentes, y se centra en el área de la conducción inteligente de vehículos, tanto en zonas urbanas como interurbanas, mediante la recogida de datos en tiempo real, medidos con sensores, por parte de los propios conductores, así como de datos capturados mediante simulación.

El objetivo de este trabajo es doble. Por un lado, el estudio y aplicación de las diferentes técnicas y métodos de detección de valores atípicos en bases de datos multivariantes, además de una comparativa entre ellos mediante las pruebas llevadas a cabo con datos de tráfico real. Y por otro lado, establecer una relación entre las situaciones anómalas de tráfico, como puedan ser atascos o accidentes, con los valores atípicos multivariantes encontrados.

La detección de valores atípicos representa una de las tareas más importantes a la hora de realizar cualquier análisis de datos, sea cual sea el dominio o área de estudio, ya que entre sus funciones primordiales se encuentra el descubrir información útil, que resulta de gran valor, y que por lo general queda oculta por la alta dimensión de los datos.

Con el uso de mecanismos de detección de valores atípicos junto con métodos de clasificación supervisada, se va a poder llevar a cabo el reconocimiento de elementos de la infraestructura vial urbana como pueden ser rotondas, pasos de cebra, cruces o semáforos.

Abstract

This work is related to the Smart Cities context, and it focuses on the area of intelligent vehicle driving, both in urban and interurban areas, through the collection of real-time sensed data by the drivers themselves, as well as data collected in a simulator.

The goal of this paper is twofold. On the one hand, the study and application of the different techniques and methods of outliers detection in multivariate databases, as well as a comparison between them through the tests carried out with real traffic data. And on the other hand, to establish a relation between anomalous traffic situations, such as traffic jams or accidents, with the multivariate outliers found.

Outliers detection represents one of the most important tasks when performing any data analysis, regardless of the domain or area of study, since among its fundamental functions is to discover useful and valuable information that usually is hidden by the high dimensionality of the data.

By means of using outliers detection mechanisms together with data classification methods, the recognition of elements of urban infrastructure such as roundabouts, zebra crossing or traffic lights will be carried out.

Índice

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN

EL trabajo realizado en esta tesis doctoral forma parte del proyecto HERMES (*Healthy and Efficient Routes in Massive Open-Data Based Smart Cities*), llevado a cabo por diferentes grupos de investigación pertenecientes a las Universidades de Vigo y A Coruña, Universidad de Sevilla y Universidad Carlos III de Madrid; en concreto, dentro del subproyecto *Smart Driving* and *Semantic Data Handling 'HERMES-SmartDriver'*, TIN2013-46801-C4-2-R, financiado por el MINECO (Ministerio de Economía, Industria y Competitividad), así como la ayuda BES-2014-070462, del programa FPI, para la realización de esta investigación.

Como se verá en los sucesivos capítulos, la labor realizada en este trabajo se centra en el análisis y tratamiento de datos, junto con la aplicación de técnicas estadísticas y de aprendizaje automático, sobre datos recogidos durante la conducción, en el marco de ciudades inteligentes. Estos datos han sido proporcionados tanto por conducciones reales, gracias al uso de la aplicación Android *SmartDriver*, como a datos recogidos mediante un entorno de simulación.

A continuación se presentan la motivación y objetivos que han dado lugar a esta tesis, seguido de una breve descripción del concepto de ciudades inteligentes, así como de la aplicación *SmartDriver*.

1.1. Motivación y Objetivos

A día de hoy, es incuestionable que la movilidad vial representa un factor clave para el desarrollo económico de cualquier comunidad, siendo una contribución

vital para el progreso de la sociedad actual y futura. Tal es así, que existen varios indicadores que refuerzan la idea de que existe una importante correlación entre la movilidad vial y el crecimiento económico. Por ejemplo, las mejoras en la movilidad reducen los costes de transporte, así como la reducción del tiempo de cada trayecto, repercutiendo en un incremento de la productividad, y por lo tanto en el crecimiento económico.

Esta tesis se encuentra inmersa en el actual paradigma de ciudad inteligente, donde el uso de las llamadas TIC, 'Tecnologías de la Información y la Comunicación', juegan un rol importante en la búsqueda de mejores soluciones en el área del transporte, de manera que se busca conseguir una movilidad que sea sostenible, y más segura y eficiente. Y es que a lo largo de los últimos años, se ha ido incrementando de forma exponencial tanto el número de vehículos, como la frecuencia de viajes, generando importantes problemas, tanto de seguridad vial, como medioambiental, de modo que es necesario afrontarlo con la aportación de nuevas soluciones que garanticen una movilidad que no solo sea segura y eficiente, sino que además sea respetuosa y sostenible con el medio ambiente.

Igualmente, se ha visto incrementado el número de accidentes de tráfico, a la par que va aumentando el parque móvil en cada ciudad o comunidad. La naturaleza de estos accidentes puede ser muy diversa, pero gracias a la implementación de nuevos sistemas de movilidad inteligente es posible que se dé una significativa mejora en el número de accidentes.

Esta movilidad inteligente puede ser lograda gracias al uso de diferentes datos medidos por varios sensores, por ejemplo aquellos que capturan la telemetría del vehículo, o los que miden la frecuencia cardíaca, proporcionando así información del estado del conductor.

El principal objetivo de este trabajo va a ser estudiar y analizar los datos recogidos durante la conducción, por parte de distintos conductores, como una variable estadística multivariante y se centrará en la detección de observaciones atípicas dentro del conjunto de datos obtenido. Estos valores atípicos servirán de base para la detección e identificación de situaciones anómalas de tráfico, como puedan ser condiciones de congestión o atascos y accidentes.

A pesar del hecho de que las técnicas de detección de atípicos tienen como fin la eliminación de tales valores, producidos por ejemplo, por errores humanos o fallos del sistema en la recogida de datos, estos valores atípicos pueden resultar

en datos anómalos pero a su vez muy útiles, ya que pueden manifestar una significante e importante información, dependiendo de la naturaleza del conjunto de datos con el que se trabaje.

Existe un gran número de aplicaciones, en muy diversas áreas, donde la detección de valores atípicos es una tarea primordial, como son los casos de detección de fraude en tarjetas de crédito, procesamiento de solicitud de préstamos, detección de intrusión en redes, monitorización de actividades, diagnóstico de fallos, detección de defectos estructurales, rendimiento en redes, análisis de imágenes de satélite, segmentación de movimientos, monitorización de series temporales, seguimiento de condiciones médicas o investigación farmacéutica, entre otras muchas (Hodge & Austin, 2.004).

El interés en detectar situaciones anómalas de tráfico, como pueden ser los atascos, surge a partir tanto de la concienciación medioambiental, como la sensibilización a los graves efectos sobre la salud provocados por la circulación viaria. La finalidad que se busca es mejorar los desplazamientos en el tráfico rodado, de modo que los vehículos puedan desplazarse de un lugar a otro de la manera más rápida y eficientemente posible, consiguiendo así minimizar los costes externos asociados con el tráfico, ya sean causados por accidentes, o por impactos ambientales adversos. Por tanto, una tarea primordial en el análisis de datos de tráfico involucraría la detección de puntos atípicos para prevenir, por ejemplo, el tiempo malgastado durante un atasco y los problemas de salud causados por la polución del aire en esa zona.

1.2. Ciudades Inteligentes

Una 'Ciudad Inteligente', o más comúnmente conocida por su término en inglés *Smart City*, es un concepto ampliamente difundido en la sociedad de hoy en día, pero cuya definición se difumina igualmente en una clara falta de estandarización respecto a lo que realmente representa. Pese a ser un término muy genérico, se podría resumir que el objetivo principal a destacar sobre una ciudad inteligente es mejorar la calidad de vida de sus ciudadanos mediante el uso de la tecnología.

Para introducir el concepto de ciudad inteligente hay que remontarse al año 2.000, donde el trabajo publicado por Robert E. Hall presenta una de las primeras aproximaciones de ciudad inteligente, definiéndola de la manera siguiente:

'Una ciudad que controla e integra las condiciones de todas sus infraestructuras críticas con el fin de optimizar mejor sus recursos, planificar sus actividades de mantenimiento, y monitorizar aspectos de seguridad, aumentando al máximo los servicios a los ciudadanos.'

También en el año 2.000, Hall et al. establecen que:

'La ciudad inteligente se convertirá en el centro urbano del futuro, realizado de forma segura, y seguro para el medio ambiente; siendo eficiente desde un punto de vista estructural en servicios como energía, agua y transporte; que estará diseñado y desarrollado haciendo uso de materiales avanzados como son sensores y electrónica, sistemas informáticos y algoritmos de toma de decisiones'.

Diez años más tarde, Donato Toppeta (2.010) la describe como:

'Ciudad que combina las TIC y la tecnología Web 2.0 con organizaciones para diseñar y planificar esfuerzos que aceleren procesos burocráticos y que ayuden a identificar soluciones innovadoras que gestionen la complejidad de la ciudad, con el fin de mejorar la sostenibilidad y calidad de vida.'

A partir de aquí la ciudad inteligente empieza a convertirse en una estrategia para ayudar a mitigar los problemas generados por el crecimiento de la población en las áreas urbanas.

En Chourabi et al. (2.012) se propone un *framework* o marco de trabajo para facilitar la comprensión del concepto de ciudad inteligente. En este marco de trabajo se identifican los factores que hay que proporcionar como base para lograr iniciativas de ciudades inteligentes. Estos factores pueden estar influenciados en mayor o menor medida por otros factores. Con el fin de reflejar niveles diferenciados de impacto, los factores están representados en dos niveles de influencia, como puede observarse en la Figura 1.1.

Los factores internos estarían representados por la tecnología, la gestión y la política, y los factores externos por las personas y comunidades, la economía, las infraestructuras, el entorno natural y el gobierno. Estos factores externos estarán de alguna manera influenciados por los factores internos, siendo la tecnología el factor que influye de manera determinante en el éxito de los demás factores del marco de trabajo.

Así en 2.014, la Unión Internacional de Telecomunicaciones introduce el término de sostenibilidad, donde el concepto de ciudad inteligente y sostenible se define como:

'Una ciudad innovadora que utiliza las TIC y otros medios para mejorar la calidad de vida, la eficiencia de operación, los servicios urbanos y la competitividad, garantizando al mismo tiempo que satisfaga las necesidades de las generaciones presentes y futuras, con respecto a aspectos económicos, sociales y medioambientales'.

Para terminar, en España, el Plan Nacional de Ciudades Inteligentes del Ministerio de Industria, Energía y Turismo (2.015), adopta la definición del Grupo Técnico de Normalización 178 de AENOR (AEN/CTN 178/SC2/GT1 N 003), donde se define a una ciudad inteligente como:

'Ciudad inteligente es la visión holística de una ciudad que aplica las TIC para la mejora de la calidad de vida y la accesibilidad de sus habitantes y asegura un desarrollo sostenible económico, social y ambiental en mejora permanente. Una ciudad inteligente permite a los ciudadanos interactuar con ella de forma multidisciplinar y se adapta en tiempo real a sus necesidades, de forma eficiente en calidad y costes, ofreciendo datos abiertos, soluciones y servicios orientados a los ciudadanos como personas, para resolver los efectos del crecimiento de las ciudades, en ámbitos públicos y privados, a través de la integración innovadora de infraestructuras con sistemas de gestión inteligente'.

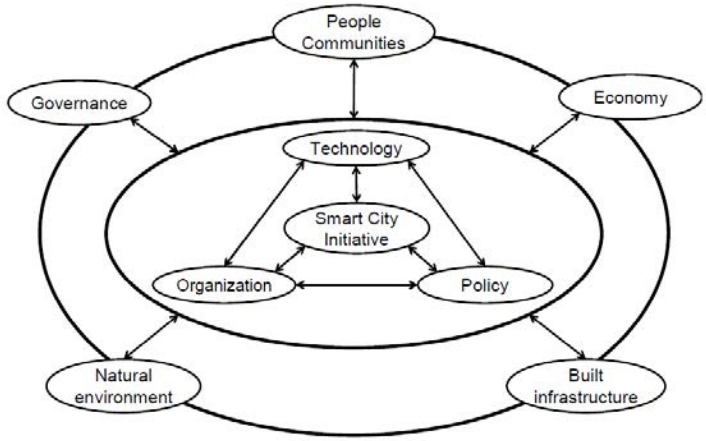

Figura 1.1 Framework para la ciudad inteligente (Chourabi et al., 2.012)

Aunque uno de los problemas básicos en el desarrollo de las ciudades inteligentes es la falta de normalización y estandarización de un modelo común de una *SmartCity*, esto no ha frenado el impulso para el desarrollo de numerosas iniciativas en ciudades de todo el mundo. No obstante, esta falta de estandarización da lugar a un desaprovechamiento de recursos y una duplicación de los desarrollos realizados.

En resumen, podemos destacar que los objetivos que se pretenden alcanzar en las ciudades inteligentes son las de convertirlas en ciudades más participativas, más eficientes, más sostenibles y cuyo fin es perseguir la mejora de la calidad de vida de sus ciudadanos, siendo la tecnología el factor que permite modificar la forma en la que se gestionan y habitan dichas ciudades. Además, gracias a los avances tecnológicos, los gestores toman mejores decisiones porque están mejor informados, los ciudadanos pueden participar activamente en el gobierno de la ciudad aportando conocimiento, y las empresas posen una mayor y mejor capacidad de impulsar el desarrollo de la ciudad.

Figura 1.2 Nube de palabras de la ciudad inteligente (ITU-T FG-SSC, 2.014)

1.3. Aplicación SmartDriver

Una de las herramientas desarrolladas dentro del proyecto HERMES es la aplicación Android *SmartDriver*. Se trata de una aplicación concebida como un asistente durante la conducción, de modo que ayude a reducir diversos factores como pueden ser el estrés, o el consumo de combustible. Para ello proporciona información sobre el conductor de forma anónima, y del estado de la vía de circulación, con el objetivo de mejorar la gestión del tráfico en las ciudades. Entre los datos recogidos durante la conducción se encuentran, entre otros, la aceleración y velocidad del vehículo. Además, por medio del GPS obtiene la posición y telemetría en tiempo real. Asimismo, la aplicación puede hacer uso de la cámara para, por ejemplo, detectar señales de tráfico. También es capaz de recoger la variabilidad del ritmo cardíaco del conductor, mediante el uso de un sensor de frecuencia cardíaca. De esta forma los vehículos actúan como sensores móviles que recogen información, que es enviada a un servidor central. El asistente estará continuamente monitorizando al conductor y enviando datos cada segundo. De igual forma que cuando detecte un valor fuera de lo normal, éste quedará automáticamente registrado y enviado al sistema central de HERMES (Corcoba y Muñoz, 2.015).

En los siguientes capítulos se describirán más en detalle todos los datos proporcionados por este asistente, ya que constituyen una de las fuentes principales para la realización de esta tesis.

En resumen, para el desarrollo de este proyecto se lleva a cabo un análisis de los datos proporcionados por la aplicación *SmartDriver,* a partir de sensores que capturan la telemetría del vehículo y el estado del conductor, con el fin de mejorar la asistencia en la conducción mediante la detección de incidencias en el tráfico viario.

Gracias a la información proporcionada por esta aplicación es posible realizar un estudio de valores atípicos en bases de datos multivariantes. Mediante la detección de estos valores atípicos se pretende identificar situaciones anómalas que suceden durante la conducción, como puede ser el caso de un atasco o accidente. Asimismo, gracias a estos atípicos se va a poder identificar la presencia de infraestructuras urbanas como, por ejemplo la entrada a una rotonda o un semáforo o un paso de cebra.

1.4. Estructura del Documento

El contenido del presente documento está estructurado en 6 capítulos. En el primero de ellos se ha presentado una breve introducción del concepto de ciudad inteligente y se han reseñado así las motivaciones para la realización de esta tesis, junto con los objetivos a alcanzar. En el capítulo 2 se repasa el estado del arte y se introduce el concepto de valor atípico, así como las diferentes técnicas desarrolladas para poder detectarlos, haciendo hincapié en el entorno de bases de datos multivariantes. También se introducen los métodos de clasificación que se han utilizado. A continuación, en el capítulo 3 se expone el modelo teórico propuesto y se repasa la metodología empleada. Además se presenta brevemente el entorno de simulación manejado. Seguido, en el capítulo 4 se detallan todos los experimentos llevados a cabo, tanto con datos de conducción en escenarios reales, como en escenarios simulados. En el capítulo 5 se exponen y discuten las conclusiones alcanzadas y los posibles trabajos futuros. Para terminar con el capítulo 6 donde se recogen las publicaciones que se han generado en el marco de esta tesis.

2. ESTADO DEL ARTE. VALORES ATÍPICOS Y CLASIFICACIÓN

EN este capítulo se repasa el estado del arte de los sistemas de monitorización de tráfico para detección de incidencias, centrándose principalmente en aquellos que usan métodos de detección de atípicos sobre datos conocidos como FCD o *Floating Car Data.* Estos datos son los proporcionados por el GPS de los teléfonos móviles de los conductores mientras se encuentran circulando.

Seguidamente, se introduce el concepto de valor atípico, no solo en variables individuales sino en un conjunto de datos multivariante, y se recoge un resumen de los métodos actuales más usados en la detección de valores atípicos multivariantes. Dado que estos atípicos no tienen por qué ser asociados con errores, es muy importante su detección, aunque no necesariamente tienen que ser eliminados del conjunto de datos. La detección de este tipo de observaciones atípicas representa una de las tareas más significativas y destacables para el descubrimiento de valiosa información, como ha sido probado en numerosas investigaciones, dando incluso lugar al descubrimiento o hallazgo de nuevos valores y relaciones.

Por último, se presentan los dos tipos de clasificadores que se han utilizado en combinación con las técnicas de detección de atípicos.

2.1. Estado del Arte

Las técnicas de monitorización de tráfico para la detección de incidencias y situaciones anómalas, así como la identificación de elementos de señalización, es una disciplina muy extendida en todo el mundo, y donde las tecnologías se pueden categorizar en tres tipos principales, que son los sistemas basados en láser, los algoritmos de reconocimiento basados en visión y los sistemas de detección basados en teléfonos inteligentes.

Dentro de los basados en láser se encuentra el modelo propuesto por Holgado-Barco, et al. (2.017), donde utilizan un sistema LIDAR (*Laser Imaging Detection and Ranging*) para obtener automáticamente un inventario geométrico de las secciones transversales de la carretera.

Ejemplos de sistemas de reconocimiento basados en visión se pueden encontrar en Mascetti et al. (2.016) y Wu et al. (2.016), donde se detectan semáforos, mediante procesamiento de imágenes en dispositivos móviles.

La detección de incidencias y situaciones anómalas de tráfico en carretera, así como la identificación de elementos de señalización en vías de circulación, es efectivamente una labor muy estudiada, en la que se han propuesto números modelos y métodos, y con diferentes enfoques, sin embargo, escasamente desde un punto de vista de detección de valores atípicos, y sobre datos recogidos por sensores durante la conducción.

Los estudios basados en la detección e investigación de valores atípicos son bastante limitados en el campo del transporte, pese a que se puede recopilar fácilmente un gran volumen de datos sobre las condiciones de tráfico, gracias a los teléfonos móviles que funcionan como sensores.

A continuación, se comentan algunos de los estudios realizados hasta la fecha, que abordan esta tarea mediante la detección de valores atípicos en datos viarios.

Zhu et al. (2.009) introducen el concepto de minería de atípicos para la detección automática de incidentes de tráfico o AID *(Automatic Incident Detection)*, basada en los datos capturados por sensores, como la posición y velocidad del vehículo, y propone un nuevo enfoque de AID en vías urbanas. En el modelo propuesto se hace una caracterización del incidente a nivel temporal y espacial, e implementa un método de detección multinivel, que consiste en un filtrado, detección de valores

atípicos y monitoreo de retardo. Para la detección de atípicos utiliza un método de detección basado en la distancia, que se explicará más adelante. Los datos han sido capturados por 13.000 taxis circulando por la ciudad de Pekín durante 10 días. Y como resultado obtienen una tasa de detección del 81,5%, calculada como el número de incidentes detectados, frente al número total de incidentes ocurridos.

También en 2.009, una aproximación llevada a cabo por Li, Han y Lee, presenta un método para detectar valores atípicos temporales, realizado con experimentos usando datos de tráfico real, también de taxis rodando durante 24 días en la ciudad de San Francisco, y con más de 33 millones de movimientos de vehículos en segmentos de carretera registrados. En este caso los datos corresponden al comportamiento conjunto del tráfico en segmentos de vía y no a datos individuales de vehículos en movimiento, como es el tema de esta tesis. Utiliza información aglomerada de todo el conjunto de datos. Para cada intervalo de tiempo compara cada segmento de carretera con otros segmentos y el historial de valores de similitud se registran en un vector local temporal en cada segmento de carretera. Los valores atípicos se calcularían a partir de los cambios drásticos en esos vectores.

En Ge et al. (2.010) se propone un método para identificar situaciones anómalas de tráfico, a partir de la evolución de las trayectorias anormales, mediante la detección de valores atípicos durante la evolución de trayectorias. Considera dos tipos de trayectorias atípicas: valores atípicos en términos de densidad y valores atípicos en términos de dirección, que son aquellos que se desvían de la mayoría de las trayectorias dentro de un espacio y tiempo observado.

Liu et al. (2.011) estudian por primera vez, según ellos, las interacciones causales que existen entre los valores atípicos detectados en datos espacio-temporales de tráfico. Proponen un algoritmo que construye árboles de causalidad atípica, basados en las propiedades temporales y espaciales de valores atípicos detectados, con datos de trayectorias recolectados por taxis en Pekín. La causalidad entre los valores atípicos implica que no solo se necesita descubrir valores atípicos del tráfico, si no también inferir relaciones causales e interacciones entre ellos. Como resultado observan interacciones recurrentes entre los valores atípicos espacio-temporales, que serán aquellos enlaces cuyos atributos espacio-temporales son muy diferentes de los enlaces vecinos. La zona urbana queda mapeada en regiones utilizando la red de carreteras, y se modela como un grafo, donde cada nodo es una región y cada enlace captura el flujo de tráfico entre dos regiones. Así consiguieron identificar en los datos de tráfico instancias reales de anomalías, como pueden ser

controles de tráfico o atascos. Además proponen otro algoritmo que puede usarse para revelar anomalías recurrentes en la red de carreteras.

Guo et al. (2.015) proponen también un sistema de detección de valores atípicos en tiempo real, con el fin de identificar eventos poco comunes que ocurren en la carretera. Utilizan datos reales de 36 estaciones de Estados Unidos y Reino Unido, y en este caso, han sido recopilados por dispositivos de vigilancia de tráfico, y sistemas de detectores de bucle inductivos, instalados en las carreteras de todo el mundo. Estos dispositivos son capaces de capturar continuamente gran cantidad de datos de condiciones de tráfico. La detección e investigación de los valores atípicos en esos datos va a proporcionar una valiosa información sobre patrones extraordinarios como accidentes de tráfico o condiciones climáticas adversas.

Su propuesta tiene un enfoque, en cuanto a la de detección de valores atípicos, basado en el modelo de variación en el tiempo de la varianza condicional de la serie de flujo de tráfico. Presenta, por tanto, un sistema de predicción del flujo de tráfico a corto plazo, que es capaz de generar conjuntamente el nivel de pronóstico y la variación de la varianza condicional, junto con un método de detección de valores atípicos. Y los efectos de los valores atípicos se investigan en el sistema de predicción, para responder de manera adaptativa a los patrones de tráfico cambiantes.

Por último, y más reciente, en la propuesta de Ma et al. (2.016), utilizan un método de detección de valores atípicos basado en la densidad, junto con análisis de componentes principales, para identificar situaciones anómalas de tráfico, ya sean accidentes o atascos, pero sobre señales grabadas en video. El conjunto de datos de flujos de tráfico empleado fue recolectado por una cámara de video en un cruce de Hong Kong, durante 31 días. Y empleando un enfoque semi-supervisado para etiquetar cualquier valor atípico, alcanzando una tasa promedio de acierto en la detección del 93.5%.

2.2. Introducción a los Valores Atípicos: Definición y Tipos

Los valores atípicos son, como su nombre indica, aquellos valores que se encuentran fuera del rango de valores esperados para un conjunto de datos dado.

Estadísticamente hablando, un valor es atípico con referencia a una observación, si en comparación a los demás datos recogidos resulta muy distante numéricamente,

esto es, un atípico es una observación extremadamente grande o pequeña, y puede tener efectos desproporcionados en los resultados estadísticos.

Estas observaciones anómalas, también llamadas datos discordantes o aberrantes, se conocen en la literatura estadística como *outliers* o *atípicos*, y en la mayoría de los casos suelen ser el resultado de un equipo mal calibrado, entrada incorrecta de datos o errores de procesamiento o codificación. También pueden aparecer por causas desconocidas o como consecuencia de una situación extraordinaria.

Se les denomina también valores influyentes dado que su presencia puede alterar notablemente los parámetros característicos de un conjunto de datos, como el valor de la media o la desviación estándar, dando lugar así a interpretaciones erróneas.

Dada la importancia de detectar estos atípicos, conviene estudiar en detalle su procedencia antes de pasar a su eliminación. No obstante, es importante resaltar el necesario interés que tiene detectar ciertos valores atípicos. Uno de los ejemplos más destacable que pone claramente de manifiesto la importancia en la detección de observaciones atípicas, lo podemos encontrar en el año 1.985, cuando varios investigadores hallaron que los datos recopilados por la *British Antarctic Survey* o Prospección Antártica Británica, mostraban que los niveles de ozono en la Antártida habían caído un 10% por debajo de los niveles normales. Sin embargo, el satélite Nimbus 7, lanzado por la NASA años antes y que tenía instrumentos a bordo para registrar niveles de ozono, no registró esas bajas concentraciones. El problema fue que las concentraciones de ozono registradas por el satélite fueron tan bajas que un programa informático las consideró como observaciones atípicas y las descartó.

Se podría decir entonces que existen dos razones esenciales para buscar y detectar valores atípicos. La primera de ellas es, obviamente debido a que los valores atípicos podrían influir en los resultados del resto de los datos, y la segunda debido al interés del valor atípico por sí mismo.

A la hora de decidir si se rechaza o no una observación extrema, según Barnett y Lewis (1.994), habría que considerar los tipos de variabilidad, de modo que si la variabilidad se debe a un error de medición o error de ejecución, esta observación debería ser eliminada. No obstante, si la variabilidad se debe a una variación inherente, debería permanecer. El problema aparece entonces cuando se desconoce la fuente de variación. Con todo, cabe puntualizar que no existe una definición formal, ampliamente aceptada, de lo que es un 'atípico' (Jollife, 2.002).

2.2.1. Atípicos Univariantes

Cuando la base de datos bajo estudio está formada por una sola variable, los atípicos, llamados univariantes, pueden encontrarse en función de varios criterios.

Si se considera que los datos, X, siguen una distribución normal, $X \sim N(\mu, \sigma)$, es fácil observar a los valores atípicos, como aquellos que se encuentran a una distancia superior de dos veces su desviación típica con respecto a su valor medio. Para dicha distribución normal, la probabilidad de encontrar un punto que está a más de tres desviaciones típicas de la media es menor de 0,0027%, y para una distribución arbitraria la probabilidad sería menor de 1/9, esto es, 0,1111%.

Probablemente el criterio más extendido es considerar como atípico aquel valor que dista más de 1,5 veces su distancia intercuartil, o incluso hasta 3 veces para un criterio más exigente, cuya representación corresponde al llamado *boxplot* o diagrama de cajas, introducido por Tukey en 1.977. Pero cuando se tiene más de una variable, este criterio ya no es aplicable.

2.2.2. Atípicos Bivariantes

En el caso de tener dos variables, los atípicos pueden identificarse fácilmente mediante un diagrama de dispersión. Un ejemplo de valores atípicos bivariante se encuentra en el grupo de estrellas *'Star Cluster CYG OB1'* (Figura 2.1.). Se trata de un conjunto de 47 estrellas en la dirección Cygnus, una de las constelaciones de la Vía Láctea, donde estas 47 observaciones forman un *data frame* de dos variables, que son el logaritmo de la temperatura efectiva en la superficie de la estrella y el logaritmo de la intensidad de la luz que emite.

En la Figura 2.1 se observa claramente como en este grupo de estrellas aparecen cuatro observaciones atípicas, correspondientes con muy altas luminosidades y con las temperaturas más bajas, que son conocidas como estrellas gigantes.

Si bien en la representación bivariante se detectan claramente estas cuatro estrellas como atípicas, queda de manifiesto en la Figura 2.2, donde se han representado los diagramas de cajas de cada una de las variables individuales, que no se pueden encontrar estos valores atípicos buscándolos individualmente en cada una de ellas.

Mientras que en la Figura 2.1 aparecen cuatro atípicos claramente distanciados del

resto de los datos, en los diagramas de cajas, la variable correspondiente a la intensidad de luz emitida no tiene ningún valor atípico y la variable de temperatura efectiva de la superficie de la estrella presenta cinco atípicos, uno de ellos con valor 3,84 y los otros cuatro quedan superpuestos en la figura y se corresponden, tres de ellos con el valor 3,49 y el último con un valor de 3,48. Esto es debido a que en el análisis bivariante hay que tener en cuenta la correlación entre ambas variables y por lo tanto los métodos para detectar las observaciones atípicas deben ser diferentes.

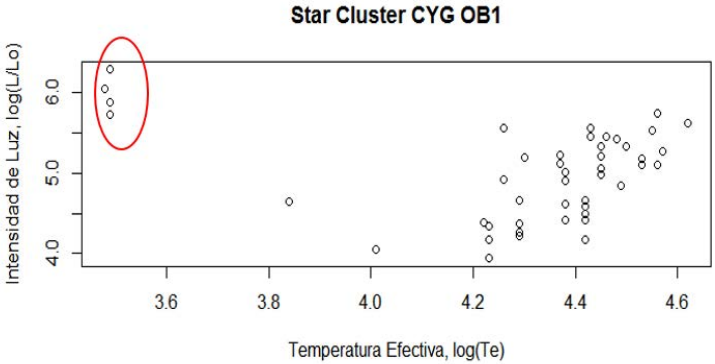

Figura 2.1 Diagrama de Hertzsprung-Russell

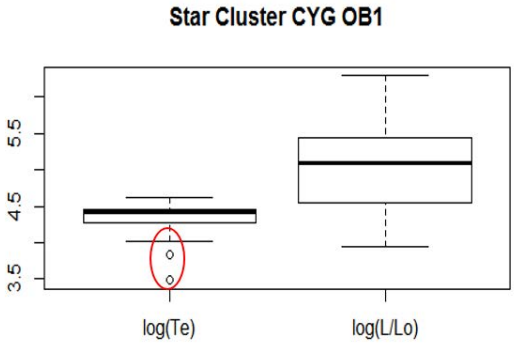

Figura 2.2 Diagrama de cajas o boxplot del cluster CYG OB1

2.2.3. Atípicos Multivariantes

Numerosos estudios, ya sea del ámbito científico, tecnológico, económico, o cualquier otra disciplina, se han enfrentado en la última década a la ardua tarea de tener que detectar observaciones atípicas en un conjunto de bases de datos multivariantes, es decir, formada por diferentes componentes o variables que forman una única variable multidimensional.

Una definición exacta de lo que se considera observación atípica depende, en gran medida, de los supuestos ocultos en relación con la estructura de los datos y el método de detección aplicado. Sin embargo, algunas definiciones se consideran lo suficientemente generales como para hacer frente a diversos tipos de datos y métodos. A continuación se recopilan por orden cronológico las definiciones más destacadas.

Grubbs (1.969) indica que una observación atípica es aquella que parece desviarse notablemente de los otros miembros de la muestra en la que ocurre.

Gnanadesikan y Kettenring (1.972), contemplan los atípicos multivariantes como observaciones que se consideran extrañas, no por el valor que toman en una determinada variable, sino en el conjunto de aquellas.

Para Rohlf (1.975), debido a la aparente complejidad del problema, se pueden caracterizar los valores atípicos por el hecho de estar aislados de la nube principal de puntos, de modo que podría ocurrir que no se observen al final de la distribución, como en el caso univariante, pero se deben manifestar de alguna forma.

Según Hawkins (1.980), un atípico es una observación que se desvía tanto del resto de observaciones como para levantar sospechas de que fue generada por un mecanismo diferente.

Johnson (1.992) define un valor atípico como una observación en un conjunto de datos que parece ser inconsistente con el resto de ese conjunto de datos.

Y Barnett y Lewis (1.994) indican que una observación periférica o atípica, es aquella que parece desviarse marcadamente de otros miembros de la muestra en la que se produce. Además, para responder a la pregunta de por qué o cómo aparecen los atípicos, los clasifican en tres grupos:

- *Variabilidad inherente*, definida como la variabilidad natural en cualquier conjunto de datos.
- *Error de medición*, esto incluye la limitación del dispositivo de medición así como cualquier error de grabación realizado por el científico.
- *Error de ejecución*, donde se incluyen situaciones con observaciones que no están en la población de interés o situaciones cuando se usa una muestra sesgada o mal calculada.

También, Beckman y Cook (1.983), hacen una clasificación de los valores atípicos distinguiendo tres tipos:

- *Observación discordante*, aquella observación que parece sorprendente o discrepante para el investigador.
- *Observación contaminante*, aquella que no proviene de la población en estudio, sino de otra población.
- *Observación influyente,* aquella que al ser excluida del análisis de datos altera sustancialmente rasgos importantes de dicho análisis.

2.3. Problemas en la Detección de Atípicos

Una primera y trivial clasificación de las técnicas de detección de valores atípicos sería en función del número de variables de que conste el conjunto de datos. Como puede verse en la Tabla 2.1, existen técnicas univariantes, que son aquellas en las que se tiene una sola variable, las técnicas bivariantes, para dos variables, y las técnicas multivariantes, cuando se tienen más de dos variables. En este último caso se puede distinguir entre multivariantes de baja dimensión y multivariantes de alta dimensión, donde pueden darse desde cientos hasta miles de variables.

En el caso de trabajar con datos multivariantes, la detección de valores atípicos presenta grandes dificultades debido, principalmente, a la dimensión del conjunto de datos. Por un lado, no se pueden considerar como valores extremos, como sucede en un entorno unidimensional, y no se pueden detectar visualmente por el problema de la dimensión, dado que un valor atípico no es necesariamente el dato más extremo en cualquiera o alguna de sus componentes.

Además una observación multivariante puede ser un valor atípico debido a un error brusco en una de sus componentes o por pequeños errores sistemáticos en

varias de ellas. Luego, en la detección de valores atípicos multivariantes se presentan varios problemas, además de que una observación puede ser identificada como atípica por un método dado y no por otro, o ser atípica en un espacio *p*-dimensional y no serlo necesariamente en un subespacio dado.

Por otra parte, es muy habitual en datos multivariantes que se den los efectos de '*masking*' o enmascaramiento y '*swamping*', de empantanamiento o inundación.

El fenómeno de enmascaramiento ocurre cuando un atípico no es declarado como tal, ya que otro valor atípico, el más cercano a él, oculta su importancia. Según Acuña (2.004) un atípico enmascara a un segundo atípico, si el segundo atípico puede ser considerado como un valor extremo sólo por sí mismo, pero no en presencia del primer atípico. Así, después de la eliminación del primer atípico, en una segunda instancia, el otro punto se convierte también en valor atípico.

Por otro lado, el fenómeno de empantanamiento ocurre cuando al aplicar una prueba de forma sucesiva para detectar atípicos múltiples, un atípico claramente discordante arrastra consigo otro valor que no necesariamente lo es. De nuevo, para Acuña (2.004), un atípico empantana una segunda observación, si esta última puede ser considerada como un valor extremo sólo bajo la presencia de la primera. Esto es, después de la eliminación del primer atípico, la segunda observación se convierte en un valor no atípico.

En otras palabras, el efecto de enmascaramiento se produce cuando un grupo de observaciones extremas sesga las estimaciones de la media y de la covarianza hacia él, y la distancia resultante del valor extremo a la media es pequeña, mientras que el efecto de empantanamiento ocurre cuando un grupo de valores extremos sesga las estimaciones de la media y de la covarianza hacia él y lejos de otros valores no periféricos, y la distancia resultante de estos casos a la media es grande, haciéndolos parecer como atípicos.

Tabla 2.1 Técnicas de detección de atípicos en función del número de variables

Tipo de Técnica	**Número de Variables**
Univariantes	Una sola variable
Bivariantes	Dos variables
Multivariantes de baja dimensión	De tres a cientos de variables
Multivariantes de alta dimensión	De cientos a miles de variables

Luego, ambos efectos conducen a detectar pocos valores atípicos o muchas veces a la imposibilidad de detectarlos. Por lo tanto, es importante señalar que la estructura de los datos juega un papel primordial para la eficiencia en la detección y búsqueda de tales atípicos.

Como no hay un orden natural en los datos multivariantes, si se quieren detectar estas observaciones atípicas es necesario imponer un cierto orden, y las medidas de distancias pueden ser utilizadas para dar dicho orden, siendo el caso más habitual el uso de la distancia de Mahalanobis. Como se verá más adelante, se trata de una medida estadística de la distancia multidimensional de un punto respecto al centroide o media de las observaciones, que se caracteriza por tener en cuenta las correlaciones entre las variables, eliminando así las redundancias que puedan existir entre éstas.

Aunque la distancia de Mahalanobis ha demostrado ser un buen detector de atípicos, hay que tener en cuenta que, por un lado, los efectos de enmascaramiento podrían disminuir esta distancia para un valor atípico dado, y por otro lado, el empantanamiento podría aumentar la distancia de Mahalanobis de las observaciones que no son atípicas. No obstante, estos problemas de enmascaramiento y empantanamiento podrían resolverse usando estimaciones robustas, las cuales, por definición, están menos afectadas por los valores atípicos (Peña, 2.002).

2.4. Métodos de Detección de Atípicos Multivariantes

Antes de entrar a repasar algunas de las técnicas más significantes existentes para la detección de valores atípicos, conviene recordar de nuevo que estos valores están formados por múltiples variables independientes, de modo que serán mucho más difíciles de detectar que en el caso univariante, dificultad que se ve incrementada con el aumento de la dimensión, ya que los atípicos pueden ser extremos en cualquier número creciente de direcciones. En una o dos dimensiones, los atípicos pueden ser fácilmente identificados con solo representar los datos gráficamente, mientras que en el caso multivariante, con más de tres dimensiones, no es posible reconocerlos de forma gráfica.

En el análisis multivariante de baja dimensión, los métodos de detección de observaciones atípicas pueden clasificarse de modo general, como refleja la Tabla

2.2, por un lado en técnicas paramétricas o estadísticas, y por otro en técnicas no paramétricas, que son aquellas basadas en la minería de datos, donde se trabaja con grandes bases de datos de altas dimensiones. La mayoría de estos métodos se basan en medidas de distancias, técnicas de agrupamiento o *clustering*, y medidas de densidad local (Ben-Gal, 2.005).

Dentro de las técnicas paramétricas se han propuesto numerosos métodos que utilizan la distancia de Mahalanobis con estimadores robustos de la media y covarianza de la distribución de datos, como Rousseeuw y Van Zomeren (1.990), Khattree y Naik (1.995) y Penny (1.996). Estos métodos se conocen como DRM (Distancia Robusta de Mahalanobis).

Existen otros métodos para la detección de atípicos basados en el estadístico de Wilks (Bacon y Fung, 1.987).

En los modelos de regresión también se han propuesto métodos que permiten detectar atípicos mediante medidas robustas (Atkinson, 1.994), e influencia residual (Barret y Gray, 1.997).

Bajo el supuesto de normalidad multivariante, Barnett y Lewis (1.994) presentan pruebas para detectar atípicos bajo los modelos de 'media desplazada' e 'inflación de la varianza'. Ambos modelos son ampliamente estudiados por Schwager y Margolin (1.982). Y Naik (1.989) construye un estadístico para la detección de atípicos en el modelo de regresión lineal multivariante utilizando resultados de los autores antes citados.

En el caso de técnicas no paramétricas, una técnica basada en minería de datos es presentada por Münz et al. (2.007), donde para detectar los valores atípicos, utilizan el algoritmo de clustering *k-means* y la distancia euclídea normalizada al centroide del grupo.

Tabla 2.2 Técnicas de detección de atípicos multivariantes de baja dimensión

Paramétricas	**No Paramétricas**
Técnicas Estadísticas	Técnicas basadas en la distancia
Estimadores Robustos	Técnicas basadas en la densidad
Análisis de Componentes Principales	Técnicas de agrupamiento *(Clustering)*
Búsqueda de Proyecciones *(Projection Pursuit)*	SVM de una clase *(One-class SVM)*

2.4.1. Técnicas Paramétricas

Las técnicas paramétricas o modelo dependientes, son aquellas que consideran que las observaciones se distribuyen según una función de probabilidad conocida. De este modo se asume que los datos presentan una distribución simétrica y con relación lineal entre las variables, luego la linealidad va a ser un supuesto implícito en todas las técnicas multivariantes basadas en medidas de correlación.

2.4.1.1. Técnicas Estadísticas

Las técnicas estadísticas son las más simples y las primeras técnicas usadas para la detección de valores atípicos unidimensionales o univariantes.

Estas técnicas asumen que los datos siguen una distribución conocida, que en la mayoría de las veces será la distribución normal, y se basan en los parámetros estimados de dicha distribución, esto es, la media y la desviación típica.

Para identificar datos atípicos, se consideran a aquellas observaciones extremas que se encuentran relativamente lejos del centro de la distribución de datos. Para calcular dichas distancias pueden emplearse varias medidas como la 'Distancia Euclídea' o la 'Distancia de Mahalanobis'.

La distancia euclídea, muy utilizada en variables unidimensionales, tiene el inconveniente de no tener en cuenta la correlación entre variables, es decir, si las variables fueran totalmente independientes, no habría ningún problema, pero si tienen algún tipo de correlación, entonces una variable influiría sobre otra, por lo que no sería adecuado su uso con datos multivariantes.

Por lo tanto, es más acertado emplear la distancia de Mahalanobis, aunque en la práctica lo que se utiliza es el cuadrado de la misma, ecuación (2-1) o MSD (*Mahalanobis Squared Distance*), también llamada simplemente distancia de Mahalanobis,

$$MD^2(x_i) = (x_i - \bar{x}_n)^t S_n^{-1} (x_i - \bar{x}_n) \qquad (2\text{-}1)$$

donde $\bar{x}_n$ representa el vector de medias para cada variable, ecuación (2-2), y S_n^{-1} es la inversa de la matriz de varianzas y covarianzas de los datos, ecuación (2-3), siendo x_i la observación cuya distancia se quiere calcular.

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2-2}$$

$$S_n = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x}_n)(x_i - \bar{x}_n)^t \tag{2-3}$$

La definición de distancia propuesta por Mahalanobis (1.936), responde a la idea intuitiva de que los puntos que se encuentran en una zona densamente poblada deberían considerarse más cercanos entre ellos que con respecto a puntos fuera de esa zona de mayor densidad.

Así, la distancia de Mahalanobis depende de los parámetros estimados de la distribución multivariante. Ésta describe la distancia entre cada punto de datos y el centro de masas, de modo que cuando un punto se encuentra en el centro de masas su distancia de Mahalanobis será cero, y cuando se encuentra distante del centro de masas su distancia será mayor. Luego, los puntos de datos que se encuentran lejos del centro de masas con mayor distancia de Mahalanobis se consideran observaciones atípicas.

Las principales ventajas de usar técnicas estadísticas es que dado el modelo probabilístico, son métodos muy eficientes y con algoritmos rápidos, de modo que el tiempo de cálculo es lineal en términos del tamaño y dimensión de los datos, de orden *O(np)*, siendo *p* el número de variables y *n* el tamaño de la muestra. Además es posible revelar el significado de los atípicos encontrados. Y también es posible detectar atípicos sin almacenar el conjunto de datos original, el cual suele ser de gran tamaño.

Aunque también presentan varios inconvenientes, ya que es difícil encontrar un modelo de distribución que se ajuste a la estructura multidimensional de los datos y requiere saber de antemano la distribución de éstos y los parámetros de dicha distribución. Asimismo, el resultado depende en gran medida del modelo o distribución utilizados, ya que un valor que es atípico en un modelo, puede no serlo en otro, y en este caso un criterio para considerar un valor como atípico podría ser que apareciera como tal, en al menos, tres distribuciones distintas. Por otra parte, mencionar que estos métodos no son adecuados para datos periódicos ni para datos categóricos.

Además, como se vio con anterioridad, los efectos de enmascaramiento pueden disminuir la distancia de Mahalanobis de una observación atípica. Y por otro lado,

los efectos de inundación pueden aumentar la distancia de Mahalanobis de observaciones que no son valores atípicos (Penny & Jolliffe, 2.001).

Pero el principal inconveniente se debe a que la media y desviación típica de la distribución son extremadamente sensibles a la presencia de atípicos, siendo un único valor atípico capaz de sesgar completamente la media. De modo que el cálculo de la distancia de Mahalanobis se va a ver afectado por tales valores atípicos.

Para resolver estos problemas se introduce el término de *'Robustification'*, que significa hacer el estimador estadístico menos sensible a los valores atípicos, es decir, hacer estimaciones robustas de los parámetros de la distribución. Por lo tanto, las distancias de Mahalanobis deben ser estimadas por un procedimiento robusto a fin de proporcionar medidas fiables para el reconocimiento de los valores extremos (Peña, 2.002), caso del que se trata a continuación.

2.4.1.2. Estimadores Robustos: MCD y MVE

Como se acaba de comentar, son una mejora con respecto a las técnicas clásicas estadísticas donde se trata de calcular estimaciones robustas de los parámetros de la distribución de datos, evitando así, entre otros, los problemas de enmascaramiento e inundación.

Los parámetros a estimar son el vector de medias $\bar{x}$, y la matriz de covarianzas *S*, sobre la muestra multidimensional, formada por n observaviones y p variables o componentes.

Para ello se parte de una submuestra de los datos, de tamaño *h*, menor o igual que el tamaño total de la muestra, *n*, obteniéndose así el vector de medias robusto $\bar{x}_R$, y la matriz de covarianzas robusta S_R.

$$h = (n + p + 1)/2 \leq n \tag{2-4}$$

Los dos estimadores más utilizados son el estimador de covarianza de mínimo determinante o MCD *(Minimum Covariance Determinant)*, y el estimador de elipsoide de volumen mínimo o MVE *(Minimum Volume Ellipsoid).*

El MCD de los datos es, por tanto, la media y la matriz de covarianza basados en la muestra de tamaño *h* que minimiza el determinante de la matriz de covarianza. Siendo *h* el número mínimo de valores que no deben ser atípicos, el MCD se define como:

$$MCD = (\bar{x}_J^*, S_J^*) \tag{2-5}$$

donde J = {conjunto de h observaciones: $\left|S_J^*\right| \leq \left|S_K^*\right|$ $\forall$ conjunto K tal que $|K| = h$ }

Y los estimadores robustos quedan definidos:

$$\bar{x}_J^* = \frac{1}{h}\sum_{i \in J} x_i \tag{2-6}$$

$$S_J^* = \frac{1}{h}\sum_{i \in J} (x_i - \bar{x}_J^*)(x_i - \bar{x}_J^*)^t \tag{2-7}$$

Dado que encontrar el MCD exacto es a menudo imposible, el algoritmo utilizado para estimar el MCD es, en cierto sentido, el estimador. Se han sugerido varios algoritmos, como por ejemplo Hawkins (1.994), que propuso un método basado en intercambiar observaciones dentro y fuera de la muestra de tamaño *h*.

En cuanto al estimador MVE, propuesto por Rousseeuw (1.984), fue el primer estimador robusto de alto punto de ruptura de localización y dispersión multivariante, y se hizo popular gracias a su alta resistencia a los atípicos, lo que lo convierte en una herramienta fiable para la detección de estos valores (Van Aelst, S. & Rousseeuw, 2.009).

Este estimador MVE queda definido como:

$$MVE = (\bar{x}_J^*, S_J^*) \tag{2-8}$$

donde J = {conjunto de h observaciones: $Vol(S_J^*) \leq Vol(S_K^*)$ $\forall$ conjunto K tal que $|K| = h$ }

El MVE busca encontrar el elipsoide de volumen mínimo que cubre un subconjunto de al menos *h* observaciones de los datos, y puede ser calculado mediante un algoritmo de remuestreo. El estimador de localización es el centro geométrico del elipsoide y el estimador de la matriz de varianza-covarianza define el elipsoide en sí, multiplicado por una constante apropiada para garantizar consistencia (Rousseeuw & Van Zomeren, 1.990).

Estos métodos tienen la ventaja de eliminar los problemas de enmascaramiento e inundación, pero algunos fallan si la fracción de atípicos es mayor que 1/(p+1),

donde *p* es la dimensión del conjunto de datos o número de variables, por lo que en grandes dimensiones una pequeña cantidad de valores atípicos puede producir estimaciones deficientes (Muñoz y Uribe, 2.013). Además se tiene que dar la condición de $n > 5p$ para evitar la llamada *Curse of Dimensionality* o 'Maldición de la Dimensionalidad' (Hubert & Debruyne, 2.010), esto es, cuanto mayor sea el número de dimensiones, más similares serán los valores de la distancia de Mahalanobis para todos los puntos.

2.4.1.3. Análisis de Componentes Principales

El Análisis de Componentes Principales, más conocido por sus siglas en inglés como PCA *(Principal Component Analysis)*, es una de las técnicas más utilizadas en todas las áreas científicas, para la extracción de características y la reducción de la dimensión, dentro de un conjunto de datos.

Se trata de una técnica estadística cuyo objetivo es la reducción de la dimensión o número de atributos, y se basa en el supuesto de que la mayor parte de la información de un conjunto de datos puede ser explicada por un número menor de variables o atributos. Luego un análisis de componentes principales tiene sentido si existen altas correlaciones entre las variables, ya que esto es indicativo de que existe información redundante y, por lo tanto, un número menor de variables serán capaces de explicar gran parte de la variabilidad total.

La selección de componentes se realiza de manera que el primero recoja la mayor proporción posible de la variabilidad original, el segundo componente debe recoger la máxima variabilidad posible no recogida por el primero, y así sucesivamente (Gironés Roig, 2.013).

En el proceso de extracción de características, el algoritmo PCA utiliza una proyección lineal mediante un producto de matrices, de tal forma que se pretende reducir el número de atributos o variables que componen el conjunto de datos, mediante una matriz de proyección lineal, *M*, que proyecte los datos, *X*, en un espacio *A* de dimensión inferior.

$$\underset{rxp}{A} = \underset{rxn}{M} \cdot \underset{nxp}{X} \tag{2-9}$$

Así pues, encuentra la proyección que minimiza el error cuadrático de los datos reconstruidos, o lo que es lo mismo, intenta maximizar la varianza en el espacio proyectado. Y estos ejes de proyección se corresponden con los primeros vectores

propios de mayor valor propio, asociado a la matriz de covarianza de los datos centrados.

El algoritmo consta de 4 pasos básicos:

1. *Sustraer la media de cada atributo o variable.*
2. *Obtener la matriz de covarianza de los datos.*
3. *Calcular los vectores y valores propios de la matriz de covarianza.*
4. *Proyección de los datos sobre los vectores principales ordenados de mayor a menor valor propio.*

En este caso, se trata de la única técnica paramétrica que no asume ninguna distribución de probabilidad en los datos y por consiguiente puede aplicarse a cualquier tipo de datos, aunque está especialmente indicado para datos normales o gaussianos.

En resumen, se trata de una transformación lineal que reduce la dimensión del conjunto de datos, mediante la proyección según la cual los datos queden mejor representados en términos de mínimos cuadrados, de manera que convierte un conjunto de observaciones posiblemente correlacionadas en un conjunto de variables sin correlación lineal llamadas 'componentes principales' (CP). Se realiza una descomposición de autovalores de la matriz de covarianza, de modo que la primera componente será una combinación lineal de las variables originales con mayor varianza, la segunda componente será una combinación lineal de las variables originales con la segunda varianza más grande, y así sucesivamente.

La idea de utilizar este método para detectar valores atípicos fue propuesta por Rao (1.964), demostrándose que resulta una técnica útil para detectar y describir posibles singularidades en los datos. Siendo más probable detectar estos atípicos, que no son evidentes a partir de las variables originales, en las últimas CP que en las primeras, y para ello Rao sugirió el test estadístico, d_{1i}^2, como la suma de cuadrados de los valores de las últimas q (<p) CP,

$$d_{1i}^2 = \sum_{k=p-q+1}^{p} z_{ik}^2 \tag{2-10}$$

donde z_{ik} es el valor de la componente principal *k* para la i-ésima observación.

Los estadísticos d_{1i}^2, con *i=1,2,.. n*, deberían ser observaciones independientes de una distribución gamma, Γ(d) , si no hay valores atípicos, de modo que una gráfica de probabilidad gamma, con un parámetro de forma, estimado adecuadamente, puede exponer los valores atípicos (Gnanadesikan & Kettenring, 1.972).

Sin embargo, pese a ser una técnica bastante adecuada para datos de altas dimensiones, no hay garantía de que funcione cuando existen grupos de atípicos, debido, sobre todo, al problema del enmascaramiento (Peña, 2.002).

2.4.1.4. Búsqueda de Proyecciones

La búsqueda de proyecciones o *Projection Pursuit* (PP), es también una de las técnicas utilizadas en el análisis de conjuntos de datos multivariantes, pero que a diferencia del PCA, no considera que los datos sigan una distribución normal, la cual queda completamente caracterizada por su vector de medias y su matriz de covarianzas.

Las técnicas de PP fueron propuestas originariamente por Kruskal (1.969), aunque la primera implementación exitosa se debe a Friedman & Tukey (1.974), quienes fueron también los que le dieron el nombre de *Projection Pursuit.*

Este método consiste en la búsqueda de índices apropiados que permitan evaluar las proyecciones de los datos en diferentes subespacios de inferior dimensión, de modo que el índice de proyección (IP) se construye de tal forma que cuantifique lo interesante de la proyección (Friedman & Tukey, 1.974).

Por tanto, si lo que se pretende es detectar observaciones atípicas multivariantes, hay que tener en cuenta que todo valor atípico multivariante es un atípico univariante cuando se proyecta la muestra en cierta dirección, de modo que interesa encontrar aquellas direcciones en las que afloren el mayor número posible de valores atípicos.

La característica más interesante de esta técnica es que es uno de los pocos métodos multivariantes capaz de evitar la llamada 'maldición de la dimensionalidad', causada por el hecho de que el espacio de alta dimensión está casi vacío, y evita este problema trabajando con proyecciones lineales de baja dimensión. (Huber, 1.985). El precio a pagar es que no es adecuado para estructuras altamente no

lineales. Por otro lado, lo más interesante de estos métodos es que son capaces de ignorar las variables irrelevantes. Aunque uno de sus mayores inconvenientes es su alto coste computacional (Huber, 1.985).

La principal ventaja de este método es que es eficaz para bases de datos multivariantes de considerable extensión, aunque tiene también como inconveniente la fuerte dependencia de los índices o criterios elegidos.

En consecuencia, esta técnica se encarga de seleccionar las proyecciones de mayor interés mediante la optimización de algún índice o criterio de proyección. Y en el caso de la detección de valores atípicos resulta de gran interés, ya que cualquier observación atípica multivariante debe aparecer como atípica, al menos en una dirección de proyección, la definida por la recta que une el centro de los datos con el dato atípico (Peña, 2.002).

Establecer o definir lo que pueda ser una proyección interesante es bastante complicado. Otros autores prefieren definir proyecciones no interesantes, como es el caso de Huber, que define la proyección menos interesante como aquella que es normal o que está muy cerca de ella.

En definitiva, esta técnica consiste en buscar índices apropiados que permitan evaluar las proyecciones de los datos en diferentes subespacios, es decir, se define un criterio de proyección y se encuentra la dirección donde ese criterio se maximiza. La técnica se encarga de buscar la dirección dónde el índice calculado sea un máximo, de modo que toda observación que al ser proyectada tenga un índice mayor que cero va a ser un posible candidato a valor atípico.

El algoritmo se puede resumir en los siguientes 6 pasos:

1. *Centrar y esferar los datos originales.*
2. *Escoger un índice.*
3. *Seleccionar una dirección.*
4. *Proyección de los datos y evaluación del índice.*
5. *Si el índice no es un máximo o un mínimo volver al paso 3.*
6. *Analizar los datos proyectados.*

2.4.2. Técnicas No Paramétricas

Las técnicas no paramétricas son todas aquellas basadas en la minería de datos o *data mining*, y tienen la ventaja respecto a las anteriores, que no es necesario que los datos se ajusten a una distribución estándar para detectar valores atípicos en entornos multivariantes.

Dentro de la minería de datos, la búsqueda y análisis de valores atípicos es una tarea esencial y muy importante, por lo que se ha introducido el concepto de minería de atípicos o *Outlier Mining*, donde se dan diferentes técnicas, ya sean univariantes o multivariantes, como las basadas en estadísticos *(Statistics-based)*, basadas en distancia *(Distance-based)*, basadas en densidad *(Density-based)*, basadas en profundidad *(Deepness-based)*, basadas en desviación *(Deviation-based)*, o basadas en agrupamiento *(Clustering-based)* (Chen et al., 2.010).

2.4.2.1. Técnicas Basadas en la Distancia

Estas técnicas fueron introducidas para contrarrestar las principales limitaciones impuestas por las técnicas estadísticas, ya que no es necesario conocer la distribución de los datos para realizar un análisis multidimensional.

Los métodos basados en distancias se basan en medidas de distancias locales, de un punto a sus vecinos, *DB(p, d)*, donde una observación *x* se considerará atípica si una fracción *p* de las observaciones del conjunto de datos está a una distancia mayor que *d* desde *x*. Este fue el primer algoritmo de detección de atípicos basado en la distancia, propuesto por Knorr & Ng (1.998).

La mayoría de las métricas existentes utilizadas para técnicas de detección de valores atípicos se definen entonces sobre los conceptos de vecindad local, como por ejemplo el algoritmo k-NN *(k-Nearest Neighbors)* o de los *k* puntos vecinos más cercanos.

En estos métodos, la distancia entre los datos puede ser computada por cualquier métrica, como la distancia euclídea, ecuación (2-11), o la distancia Manhattan, ecuación (2-12).

$$D(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (2\text{-}11)$$

$$D(x,y) = \|x - y\| = \sum_{i=1}^{n} |x_i - y_i| \tag{2-12}$$

La principal ventaja de estos métodos es que son bastante sencillos y fáciles de entender y aplicar. Al mismo tiempo que son más eficientes en su cálculo que los métodos estadísticos, teniendo un tiempo de cálculo del orden $O(p \cdot n^2)$, siendo p el número de variables y n el tamaño de la muestra.

Asimismo, estos métodos se adaptan mejor al espacio multidimensional y pueden calcularse mucho más eficientemente que los métodos estadísticos. Sin embargo, adolecen de varios inconvenientes como la dependencia de los parámetros p y d, además de carecer de una buena escalabilidad para conjuntos de datos de gran tamaño, no siendo eficaces para altas dimensiones debido a la maldición de la dimensionalidad (Zhang, 2.013).

2.4.2.2. Técnicas Basadas en la Densidad

Las técnicas basadas en la densidad se basan en agrupar los datos en regiones con densidades similares, las cuales se separan de otras regiones que tienen distintas densidades. No solo implica conocer la densidad local del punto observado, sino también las densidades locales de sus vecinos más cercanos. La medida de atipicidad de un punto de datos es relativa en el sentido de que es una relación de densidad de ese punto en función de las densidades promediadas de sus vecinos más próximos (Zhang, 2.013).

El primer método de este tipo creado específicamente para la detección de valores atípicos fue desarrollado por Breunig et al. (2.000), donde se mide el grado de un objeto, posible atípico, con respecto a la densidad de la vecindad local, obteniendo así atípicos locales. Este grado de atipicidad se conoce como LOF (*Local Outlier Factor*), e intenta cuantificar cómo de alejado está el atípico.

El LOF de un objeto (ecuación 2-13) se basa en el parámetro MinPts, que es el número de vecinos más cercanos utilizados para definir el vecindario local del objeto, y contra cuyos integrantes se van a realizar las mediciones para determinar el valor de atípico. Luego el valor LOF de un objeto p representa el grado en que p es un atípico, y se calcula como la media de los coeficientes de la densidad local de accesibilidad, o *lrd - local reachability density* (ecuación 2-14) de p y los de puntos vecinos más cercanos.

La densidad local de accesibilidad de *p* es el inverso de la distancia media entre *p* y los objetos en su vecindad. Y $|N_{MinPts}(p)|$ es el número de puntos que caen dentro del vecindario *MinPts* de *p*.

$$LOF_{MinPts}(p) = \frac{\sum_{o \in N_{MinPts}(p)} \frac{lrd_{MinPts}(o)}{lrd_{MinPts}(p)}}{|N_{MinPts}(p)|} \tag{2-13}$$

$$lrd_{MinPts}(p) = 1/\left(\frac{\sum_{o \in N_{MinPts}(p)} reach - dist_{MinPts}(p,o)}{|N_{MinPts}(p)|} \right) \tag{2-14}$$

La distancia accesible del punto *p* con respecto al objeto *o* es definida como:

$$reach - dist_{MinPts}(p,o) = \max(MinPts\_dist(o), dist(p,o)) \tag{2-15}$$

El algoritmo LOF compara entonces la densidad local de un punto con la de sus vecinos, de modo que si:

- LOF ≈ 1 -> Densidad local del punto es parecida a la de sus vecinos.
  El objeto está dentro del cluster. No debe ser atípico.
- LOF > 1 -> Densidad local del punto es inferior que la de sus vecinos.
  Posible valor atípico.

En la Figura 2.3 se muestra un ejemplo de dos grupos, *C1* y *C2* y dos valores atípicos, o_1 y o_2, en el sentido de densidad local, mientras que en marco de valores atípicos basados en la distancia, sólo el punto o_1 sería considerado atípico.

Por tanto, como ventaja principal cabe destacar que por lo general son algoritmos más robustos y eficaces que los basados en la distancia. Pero tiene varios inconvenientes ya que los algoritmos son más complejos que los anteriores y más costosos computacionalmente. Además no pueden manejar flujos de datos de manera eficiente, y tienen una dependencia total del parámetro *MinPts.*

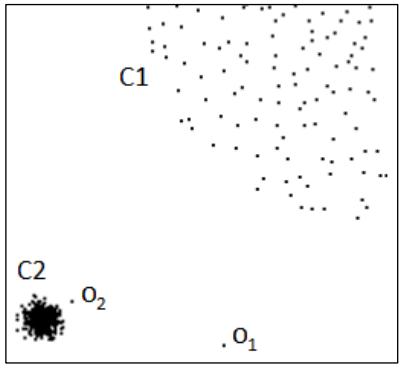

Figura 2.3 Ejemplo LOF con dos valores atípicos (Breunig et al., 2.000)

2.4.2.3. Técnicas de Agrupamiento o Clustering

El objetivo de estos métodos es formar agrupaciones de observaciones, de modo que se maximice tanto la similitud intragrupos como la diferencia intergrupos.

El *clustering* es una técnica de aprendizaje no supervisado en el cual los datos se agrupan de acuerdo a características similares, y se lleva a cabo mediante algoritmos de agrupación iterativos. Se considera que cuanto mayor sea la distancia entre una observación con el resto de los datos, mayor es la posibilidad de considerar dicha observación como atípica, de modo que si se encuentra un grupo de pequeño tamaño se puede considerar como valores atípicos agrupados.

El algoritmo más utilizado es el *K-Means,* método de agrupamiento en *k* grupos, donde un valor pertenece al grupo cuyo centroide está más cerca. Iterativamente, se van actualizando los centroides en función de las asignaciones de puntos a los *clusters* o grupos, hasta que los centroides dejen de cambiar.

Otro algoritmo sería el *K-Medoids*, donde utiliza medianas en vez de medias para limitar así la influencia de los atípicos. Algunos ejemplos son los algoritmos PAM (*Partitioning Around Medoids*) y CLARA (*Clustering Large Applications*).

Aunque estas técnicas son bastante intuitivas y consistentes con la percepción humana, estos métodos no siempre están optimizados para la detección de atípicos, ya que su principal objetivo es la agrupación, por tanto, los criterios de detección de atípicos están implícitos y no puede fácilmente inferirse de los procedimientos de agrupamiento. De igual forma, debe tenerse en cuenta que el algoritmo *k-means* es

extremadamente sensible a los valores atípicos, teniendo un impacto desproporcionado en la configuración del grupo final. Esto puede dar lugar a muchos falsos negativos, es decir, datos que deben declararse atípicos están enmascarados por el agrupamiento, y también falsos positivos con datos que están incorrectamente etiquetados como valores atípicos (Chawla, S. and Gionis, 2.013).

Otro inconveniente de estos algoritmos es que se debe conocer a priori el número de grupos, *k.* Además solo serán adecuados si el número de atípicos es pequeño.

2.4.2.4. SVM de una clase o One-class Support Vector Machine (OCSVM)

Las máquinas de vectores soporte o SVM (*Support Vector Machines*) son un tipo de clasificador binario, que pueden usarse además como máquinas de regresión y para la detección de datos nuevos o desconocidos y atípicos. Entran dentro de la categoría de métodos basados en kernel, los cuales son aplicados tanto a datos supervisados como no supervisados.

Los algoritmos de clasificación de una clase intentan encontrar el soporte de una distribución que sea capaz de clasificar automáticamente los puntos como atípicos en una gran cantidad de datos.

Una SVM de una clase, u OCSVM, usa una función de transformación implícita definida por el kernel, para proyectar los datos en un espacio de mayor dimensión. El algoritmo aprende entonces el límite de decisión, un hiperplano, que separa la mayoría de los datos del origen, de modo que los valores atípicos serían los puntos de datos que caen al otro lado del límite de decisión (Amer et al., 2.013), es decir, se busca el hiperplano que tenga la máxima distancia con los puntos que estén más cerca de él mismo.

Por consiguiente, la OCSVM es un método de detección de atípicos no supervisado, que no asume ninguna forma paramétrica de la distribución de los datos, pero es capaz de capturar la estructura real de los datos, y que además funciona mejor cuando los datos no se distribuyen como una normal. Aunque, estrictamente hablando, la OCSVM no es un método de detección de valores atípicos, sino un método de detección de datos nuevos, donde el conjunto de entrenamiento no debe estar contaminado por atípicos (Schölkopf, 1.999).

2.5. Clasificadores

Junto con los métodos de detección de valores atípicos, se han empleado métodos supervisados de clasificación, siendo el modelo Logit y las SVM los clasificadores utilizados, por ser considerados de los más eficientes. En este sentido, no se han utilizado técnicas más modernas de clasificación basadas en aprendizaje profundo, pues requieren de conjuntos de datos muy grandes para su entrenamiento, que han quedado fuera del alcance del presente trabajo.

2.5.1. Logit

El modelo Logit es un modelo de regresión logística binaria. Se trata de una técnica estadística predictiva muy simple, de una variable dependiente binaria y, variable que sólo puede tomar dos valores, en función de varias variables independientes, que pueden ser cuantitativas o cualitativas.

El modelo Logit se define como:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n \tag{2-16}$$

donde p representa la probabilidad de que ocurra el evento, $p[y=1]$, x_i son las variables independientes, β_0 la ordenada en el origen y β_i los coeficientes asociados a cada variable, los cuales son estimados mediante el método de máxima verosimilitud.

Y la función de distribución logística vendrá dada por:

$$p[y=1] = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_n x_n}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_n x_n}} \tag{2-17}$$

Dado que la variable de salida solo puede tomar dos valores, se tiene un modelo de clasificación binaria que proporciona directamente la probabilidad de pertenecer a cada una de las clases.

Se supone que la probabilidad condicional de una clase es igual a una combinación lineal de las variables de entrada, transformadas por la función logística.

El modelo de regresión logística es un modelo de regresión no lineal, aunque sí es lineal en escala logarítmica atendiendo a su definición original. En este tipo

de modelos no es posible estimar directamente los parámetros β_i, ya que son modelos no lineales. Lo que se tiene en cuenta es el signo de los estimadores, o coeficiente $\beta = \beta_0 + \beta_1 x_1 + \ + \beta_n x_n$.

Se define entonces, como cociente de probabilidades o ratio de riesgo, OR *(Odds Ratio):*

$$OR = \frac{p}{1-p} = \frac{p[y=1]}{p[y=0]} = e^{\beta} \tag{2-18}$$

De manera que el OR evalúa la influencia que cada variable independiente tiene sobre la salida. Un coeficiente β positivo significa que incrementos en la variable asociada causan incrementos en p, aunque se desconoce la magnitud de los mismos, y por tanto aumenta el OR, es decir, aumenta la probabilidad de que se de ese suceso, mientras que por el contrario, si se tiene un coeficiente β negativo disminuye el OR, esto es, un incremento en la variable asociada causará una disminución en p. Si se tiene un valor OR = 1 indicará que la probabilidad de pertenecer a una clase u otra es la misma.

El modelo Logit, en la ecuación (2-16), quedará definido en función del OR como:

$$Logit = \ln(OR) \tag{2-19}$$

Este modelo simple se cumple para variables normales, pero también para otras distribuciones, y conduce a una buena regla de clasificación, lo cual ha llevado a que sea uno de los clasificadores binarios más utilizados.

2.5.2. SVM

Las SVM, o *Support Vector Machines*, como se comentó en el apartado anterior, son un método supervisado de clasificación, aunque también pueden utilizarse en regresión o en la detección de atípicos.

El método SVM, partiendo de un conjunto de datos etiquetados en diferentes clases, usa un mapeo no lineal, Φ, transformando los datos originales en otros de dimensionalidad superior. En este espacio trata de separar las diferentes clases mediante un hiperplano óptimo, que tiene la máxima distancia con los puntos que estén más cerca de él mismo. Parte de la idea de dividir de forma

lineal un conjunto de múltiples dimensiones, donde se crean muchos hiperplanos que dividen las observaciones.

Se trata por tanto de un clasificador de margen máximo, donde los datos etiquetados con una categoría o clase estarán a un lado del hiperplano y los casos que se encuentren en la otra categoría estarán al otro lado, como puede verse en la Figura 2.4.

Estas técnicas para clasificación binaria fueron desarrolladas por Cortes & Vapnik (1.995).

La Figura 2.4 muestra un ejemplo donde se puede ver el hiperplano óptimo de separación entre las dos clases, y los puntos que caen sobre los límites, conocidos como vectores soporte.

La SVM es un algoritmo que, a partir del producto escalar de dos vectores multidimensionales, busca hiperplanos que separen los grupos. La función que define este producto escalar la denominaremos kernel.

Para clasificación, el SVM se plantea como un problema de optimización en el que se busca maximizar la distancia entre categorías, sujeto a un coste y a un número óptimo de patrones de entrenamiento.

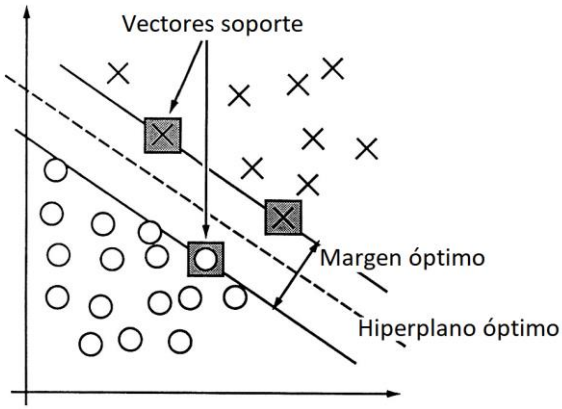

Figura 2.4 Ejemplo de clasificación SVM, caso lineal (Cortes & Vapnik, 1.995)

La simplicidad del algoritmo proviene del hecho de que SVM aplica a los datos un método simple, pero en un espacio de características de alta dimensión no linealmente relacionado con el espacio de entrada y no implica ningún cálculo en ese espacio de alta dimensión.

Las SVM utilizan un mapeo implícito , Φ: X → H , de los datos de entrada en un espacio de características de alta dimensión, definido por una función kernel, es decir, una función que devuelve el producto interno, $\langle \Phi(x), \Phi x' \rangle$, entre las imágenes de dos puntos de datos x, x' en el espacio de características.

El aprendizaje tiene lugar en el espacio de características, y los puntos de datos solo aparecen dentro de productos de puntos con otros puntos. El producto interno puede ser representado por una función kernel k:

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle \tag{2-20}$$

que es computacionalmente más simple que proyectar explícitamente x y x' en el espacio de características H.

Una ventaja de las SVM es que pueden emplearse para datos que no son linealmente separables, gracias a la utilización de distintos tipos de kernel que pueden ser, además del lineal, polinómico, radial normal (RBF - G*aussian Radial Basis Function),* Laplace RBF, función Bessel y sigmoidal.

Los kernels RBF gaussiano y Laplace, y Bessel son de propósito general usados cuando no hay conocimiento previo de los datos. El kernel lineal es útil cuando se trabaja con grandes vectores de datos dispersos, como suele ser el caso en la categorización de texto, mientras que el kernel polinomial es más usado en el procesamiento de imágenes y el kernel sigmoide se utiliza principalmente como un proxy para redes neuronales (Karatzoglou et al., 2.006).

Una propiedad interesante de las máquinas de vectores de soporte y otros sistemas basados en kernel es que, una vez que se ha seleccionado una función de kernel válida, se puede trabajar prácticamente en espacios de cualquier dimensión sin ningún coste computacional adicional significativo, ya que el mapeo de características nunca es efectivamente realizado.

2.6. Notas Finales

En relación a los valores atípicos, es importante tener en cuenta que no hay un método o técnica que sea mejor que otro, ya que la elección de éste depende de muchos factores tales como la estructura del conjunto de datos, la dimensión y tamaño de la base de datos, el tipo de datos, si es una distribución multivariante normal o no, el tipo y la proporción de valores atípicos, si se requiere una detección en tiempo real o no,… Además, una observación puede ser identificada como atípica por un método, pero no por otro. Y una observación que sea atípica en un espacio *p*-dimensional puede no serlo en un subespacio dado. Luego una solución al problema general de detección de atípicos en datos multivariantes sería la combinación de distintos algoritmos.

Tal propuesta es la presentada en el siguiente capítulo, donde se ha utilizado el algoritmo propuesto por Peña y Prieto (2.001), donde se combinan las técnicas de búsqueda de proyecciones y estimación robusta de la distancia de Mahalanobis.

En cuanto a la clasificación, mencionar que los modelos de regresión lineal son óptimos para clasificar variables que se distribuyen como una normal multivariante, mientras que pueden funcionar mal cuando se alejan de la distribución normal. Aun así, es una buena regla de clasificación binaria, ya que maximiza la separación entre ambas clases, independientemente de cual sea la distribución de los datos.

Además, tanto el modelo Logit como las SVM pueden utilizarse también en problemas de clasificación de más de dos clases.

3. PROPUESTA DE MARCO TEÓRICO METODOLOGÍA Y ANÁLISIS DE DATOS

EN este capítulo se presenta el modelo planteado y las aportaciones teóricas necesarias para resolver el objetivo de reconocimiento de situaciones anómalas de tráfico, gracias al uso de técnicas de detección de atípicos y clasificadores. Se van a describir las fuentes de información con las que se ha trabajado, tanto con sistemas reales como en un entorno de simulación, y se hará un repaso de las variables obtenidas para la generación de la estructura de datos multidimensional. Además, se presenta el algoritmo propuesto por Peña y Prieto (2.001) para la detección de atípicos multivariantes.

3.1. Marco Teórico Propuesto

El objetivo principal de esta tesis, como ya se ha mencionado con anterioridad, intenta resolver el problema del reconocimiento de situaciones irregulares de tráfico, como son los atascos o accidentes, así como la detección de elementos viarios, como puedan ser intersecciones o rotondas, mediante la detección de valores atípicos, en combinación con técnicas de clasificación, sobre datos generados durante la conducción.

Se trata de un mecanismo automático, donde solo se utilizan los datos provenientes del GPS de un dispositivo móvil, junto con un sensor de frecuencia cardíaca. Así,

de esta manera se consigue minimizar los requisitos del sistema y simplificar la recopilación de datos de muchos conductores, ocasionándoles un mínimo impacto.

La detección de valores atípicos se va a realizar, por un lado en una variable multivariante generada a partir de los datos proporcionados por la aplicación *SmartDriver*, que son capturados a través de la señal GPS, además de un sensor de frecuencia cardíaca. Por otro lado, se realizará una extracción de atípicos en la variable unidimensional correspondiente a la aceleración del vehículo.

El modelo propuesto lleva a cabo, en primer lugar, una detección de valores atípicos sobre los datos recogidos durante la conducción, empleando varios algoritmos, y a partir de éstos se utilizarán distintas técnicas de clasificación supervisada, en diferentes pruebas que serán abordadas en el siguiente capítulo.

Se trata así pues, de un enfoque novedoso en el uso de la detección de atípicos, tanto multivariantes como univariantes, no tratado antes en esta área, y cuyo modelo se presenta en la Figura 3.1.

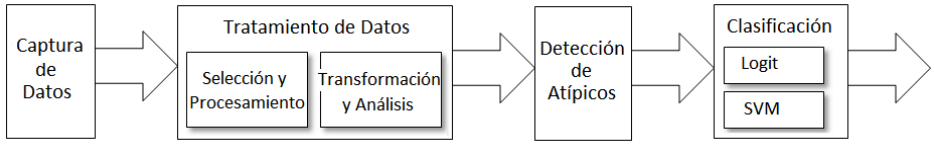

Figura 3.1 Etapas del modelo propuesto

3.2. Metodología

Una vez efectuado el proceso de adquisición de datos, es necesario realizar un tratamiento de éstos, que incluye la selección, limpieza, enriquecimiento, reducción y transformación de los datos recibidos en crudo, y que se llevará a cabo con las distintas bases de datos utilizadas.

Después de procesados los datos, se va a generar la variable multivariante sobre la que se procederá a la detección de valores atípicos.

A continuación, se menciona brevemente el concepto de entorno multivariante, seguido de la descripción de la base de datos proporcionada por la aplicación *SmartDriver* y cómo se va a crear esta variable multivariante que describe el comportamiento de las conducciones realizadas.

3.2.1. Entorno Multivariante

Bajo el nombre de análisis multivariante, como ya se adelantó en el capítulo anterior, se agrupan diversas técnicas estadísticas, dónde se estudian simultáneamente más de dos variables, y cuyo denominador común subyace en la búsqueda de una reducción de la dimensión inicial del problema. Luego el objetivo principal de las técnicas multivariantes no es la resolución de un problema estadístico, sino su simplificación, representando los datos en un espacio de menor dimensión y con la menor pérdida de información posible (Montanero, 2.008). Y gracias a estas simplificaciones es posible descubrir particularidades en los datos que son imposibles de revelar por la complejidad inicial de los mismos.

No obstante, uno de los principales problemas del análisis multivariante es que, como en todo enfoque estadístico, si bien se puede aceptar que una variable aleatoria se ajusta a un modelo de distribución normal, resulta bastante difícil aceptar la normalidad multivariante de cualquier vector aleatorio de dimensión *p*.

Esta información de carácter multivariante va a quedar representada por una matriz de datos X_{nxp}, donde *n* representa el número de observaciones y *p* las distintas variables, y será el punto de partida de los algoritmos aplicados en nuestro modelo.

$$X_{nxp} = \begin{bmatrix} x_{11} & \cdot & \cdot & \cdot & x_{1p} \\ \cdot & \cdot & & & \cdot \\ \cdot & & \cdot & & \cdot \\ \cdot & & & \cdot & \cdot \\ x_{n1} & \cdot & \cdot & \cdot & x_{np} \end{bmatrix} \tag{3-1}$$

3.2.2. Base de Datos Empleada

En las diferentes pruebas realizadas se han utilizado varias bases de datos. La principal es proporcionada por la aplicación *SmartDriver*, donde los datos son capturados por varios sensores, el GPS del teléfono móvil y un sensor de frecuencia cardíaca Polar H7, consistente en una banda conectada por bluetooth a la aplicación.

Dichos datos son enviados en tiempo real a un centro de control, donde son recopilados y tratados, y se corresponden con un total de 25 variables, formadas

tanto por datos cuantitativos como cualitativos y que se detallan en la Tabla 3.1.

Estas variables, en función del tipo de evento, pueden ser:

- Observaciones periódicas de la posición y velocidad del vehículo, generadas cada segundo, así como el tiempo entre los dos últimos latidos del corazón del conductor. Estos eventos se corresponden con *Vehicle Location, Vehicle Speed* y *RR*, respectivamente.
- Observaciones asociadas a un tramo de conducción de 500 metros. Se trata de la frecuencia cardíaca (*Heart Rate)* y la velocidad del vehículo (*Vehicle Speed*), transmitiéndose la media y desviación típica de dichos eventos. Además de la aceleración positiva de la energía cinética (*Pke*).
- Observaciones provocadas por un evento puntual. Se trata de las observaciones *High Acceleration, High Deceleration, High Speed* y *High Heart Rate,* las cuales son generadas cuando se exceden los valores normales de aceleración, desaceleración, velocidad y frecuencia cardíaca, respectivamente.

3.2.3. Generación de la Variable Multivariante

En primer lugar, en la Tabla 3.2 se muestran los distintos eventos recogidos con sus respectivas unidades. A partir de los datos en crudo y tras el procesamiento requerido, se van a considerar solo las observaciones periódicas generadas cada segundo, es decir, la posición del vehículo y su velocidad, así como el valor de RR, el cual se detallará más adelante, así como el resto de variables presentadas a continuación:

Las variables con las que se va a trabajar, obtenidas de los eventos mencionados son:

- *Velocidad media*
- *PKE*
- *Velocidad instantánea*
- *Aceleración media*
- *RR*
- *pNN50*

Tabla 3.1 Formato de los datos SmartDriver

Columna	Dato	Descripción
1	Timestamp de la observación	Número de segundos transcurridos desde el 1 de enero de 1970
2	Tipo de observación	High Acceleration, High Deceleration, High Heart Rate, Heart Rate, High Speed, Vehicle Speed, Vehicle Location, Pke y RR
3	Identificador del conductor	Cadena alfanumérica de 64 caracteres
4	Timestamp de la observación	Asociado al final de un tramo
5	Latitud del vehículo	Asociado al final del tramo
6	Longitud del vehículo	Asociado al final del tramo
7	Exactitud en metros	Medida de posición de las columnas 5 y 6
8	Velocidad instantánea del vehículo	Medida en km/h
9	Timestamp	Asociado al inicio de un tramo
10	Latitud	Al inicio del tramo
11	Longitud	Al inicio del tramo
12	Exactitud en metros	Medida de posición de las columnas 10 y 11
13	Valor de la observación	High Acceleration, High Heart Rate , High Deceleration, High Speed, Pke
14	Valor medio en un tramo	Para Heart Rate y Vehicle Speed
15	Valor mediano en tramo	Para Vehicle Speed
16	Desviación típica en tramo	Para Heart Rate y Vehicle Speed
17	Valor mínimo en tramo	Para Vehicle Speed
18	Valor máximo en tramo	Para Vehicle Speed
19	Valor de RR	Para Vehicle Location
20	Identificador del segmento de vía en OpenStreetMap	Número entero
21	Velocidad máxima	Establecida para ese segmento de vía
22	Nombre de la vía	Si se conoce
23	Tipo de vía	highway,highway_link, trunk, trunk_link, primary, primary_link, secondary, secondary_link, tertiary, tertiary_link, residential, road, unclassified, service, living_street, pedestrian, track, path, cicleway, footway, steps
24	Longitud segmento de vía	Medida en metros
25	Posición del vehículo en ese segmento de vía	Valor entre 0 y 1, donde 0 es al inicio del mismo, 1 es al final, 0.5 justo en el medio

Por un lado se van a calcular la velocidad y aceleración media en cada intervalo considerado, así como la velocidad en el instante final del intervalo, y por otro lado se obtendrán los valores de PKE, RR y pNN50.

Dado que estas variables presentan diferentes unidades, es necesario realizar una estandarización de los datos, a fin de normalizar las diferentes escalas y unidades, y poder así trabajar con ellas para llevar a cabo la detección de valores atípicos.

Tabla 3.2 Tipos de eventos y sus unidades

Tipo de Evento	**Unidades**
Velocidad del Vehículo *(Vehicle Speed)*	km/h
Frecuencia Cardíaca *(Heart Rate)*	pulsaciones/minuto
Aceleración Alta *(High Acceleration)*	m/s^2
Desaceleración Alta *(High Deceleration)*	m/s^2
Frecuencia Cardíaca Alta *(High Heart Rate)*	pulsaciones/minuto
Velocidad Alta *(High Speed)*	km/h
PKE	m/s^2
RR	ms

3.2.3.1. PKE

La variable PKE o *Positive acceleration Kinetic Energy per distance*, es la aceleración positiva de la energía cinética, definida por Watson et al. (1.985), Figura 3.2 y descrita en la ecuación (3-2).

El término PKE representa la suma de los cambios de la energía cinética positiva en un trayecto durante una maniobra de aceleración. Se expresa en m/s^2 y representa el trabajo realizado por unidad de distancia.

Se trata, pues, de un indicador que representa la habilidad de mantener la energía cinética de un vehículo tan baja como sea posible, de manera que un conductor nervioso, que presenta patrones de conducción con grandes y frecuentes fluctuaciones de velocidad, estaría asociado con un alto valor de PKE, mientras que por el contrario una conducción suave vendría asociada con un valor de PKE próximo a cero (Andrieu and Saint Pierre, 2.012).

Un ejemplo del cálculo de PKE se muestra en la Figura 3.2, siendo su definición formal la presentada en la siguiente ecuación:

$$PKE = \frac{\sum\left(V_f^2 - V_i^2\right)}{d}, \frac{dv}{dt} > 0 \qquad (3\text{-}2)$$

donde V_f y V_i representan las velocidades final e inicial, respectivamente, medidas en metros por segundo (m/s), durante el intervalo de tiempo en el cual la aceleración del vehículo es positiva, y d es la distancia total en metros, recorrida en ese intervalo.

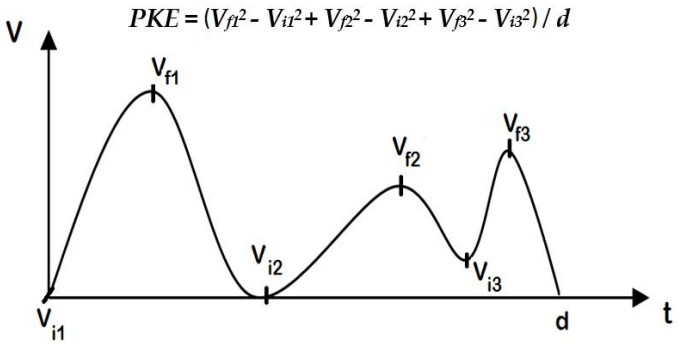

Figura 3.2 Cálculo de PKE

Este valor de PKE es ampliamente utilizado como un indicador de la congestión de tráfico. Al incrementarse solo durante las maniobras de aceleración positiva, debido a la forma cuadrática de la ecuación (3-2), se va a incrementar sustancialmente más con las aceleraciones fuertes que con las aceleraciones débiles.

3.2.3.2. RR

Un electrocardiograma usual consta de cinco ondas, P, Q, R, S y T (se suele omitir la onda final U que es de escaso valor), como puede verse en la Figura 3.3. La variable RR será el intervalo de tiempo que se corresponde con la distancia entre dos ondas R consecutivas del electrocardiograma, y que es equivalente al tiempo entre dos latidos del corazón.

El rango normal de valores para este intervalo oscila entre 0,8 y 1 segundo

(Hampton, 2.013), como queda reflejado en la Figura 3.4, lo que se corresponde con una frecuencia cardíaca de 60 a 100 pulsaciones por minuto (ecuación 3-3), donde HR o *Heart Rate* representa el ritmo cardíaco.

$$RR = \frac{60}{HR} \tag{3-3}$$

Gracias a este parámetro se va a poder obtener la variabilidad de la frecuencia cardíaca o HRV (*Heart Rate Variability*), medida que se suele emplear para determinar el nivel de estrés del conductor.

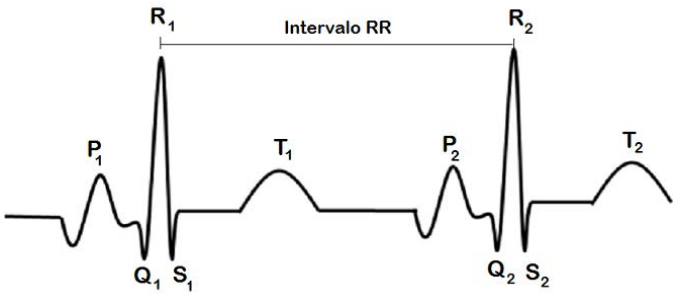

Figura 3.3 ECG con dos latidos del corazón

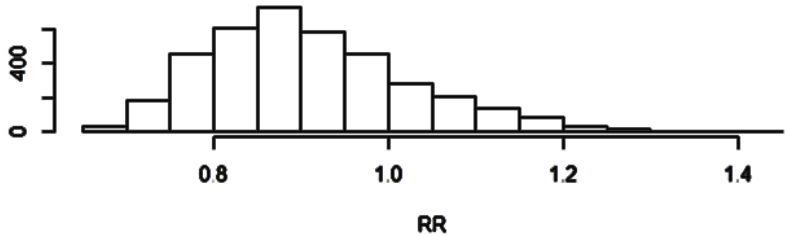

Figura 3.4 Histograma de los valores de RR recogidos durante 15 días en un tramo de autovía de unos 5 km, en Madrid

3.2.3.3. pNN50

El estadístico pNN50 es una medida de la variabilidad de la frecuencia cardíaca o HRV, en el dominio del tiempo, y se trata de una de las variables derivadas de la diferencia entre los intervalos RR.

La HRV es la variación del intervalo entre dos latidos del corazón, que puede obtenerse a través del ritmo cardíaco, y cuyo aparente fácil cálculo ha popularizado el uso de esta medida, ya que resulta ser una variable muy útil al permitir detectar el nivel de estrés del conductor, de modo que un estrés elevado queda reflejado por un alto valor de HRV, mientras que un valor bajo de HRV indicaría un nivel bajo de estrés.

Como se puede ver en la ecuación (3-5), la variable pNN50 representa el porcentaje del valor NN50, que es el número medio de veces que los intervalos RR adyacentes varían en más de 50 milisegundos (Miteus et al., 2.002).

En la ecuación (3-4) se muestra el cálculo de la variable NN50, donde NN_i representa el actual intervalo NN (Normal a Normal o RR), NN_{i-1} es el intervalo previo y k es el número total de intervalos.

$$NN50 = \sum_{i=1}^{k} \left[\left(NN_i - NN_{i-1} \right) > 50ms \right] \tag{3-4}$$

$$pNN50 = \frac{NN50}{k} \cdot 100 \tag{3-5}$$

Con el cálculo de estas variables, junto con la velocidad y la aceleración, se obtendrá la variable multivariante o matriz de datos X, que es el punto de partida de nuestro modelo.

Otras pruebas han sido realizadas con una base de datos creada bajo un entorno de simulación, que se presenta brevemente a continuación.

3.2.4. Etapas del Proceso de Clasificación

Las etapas para llevar a cabo la clasificación de una determinada observación, o vector de entrada, y que sea asignada a una categoría o clase, puede verse en la Figura 3.5.

En primer lugar es necesario seleccionar un vector de características, las cuales deben tener una buena capacidad de discriminación, de manera que los valores medidos se mantengan muy similares entre observaciones de la misma clase, y diferentes entre observaciones de distinta clase.

En segundo lugar, los datos tienen que dividirse en dos partes, una parte de entrenamiento y otra de prueba o test, para la validación del modelo de clasificación.

En nuestro caso, para lograr una mayor cantidad de datos de entrenamiento, se ha realizado una validación cruzada de *k* iteraciones, o *k-fold cross validation*, donde aleatoriamente se divide el conjunto de entrenamiento en *k* conjuntos disjuntos de igual tamaño. Y el porcentaje de aciertos estimado será la media de las *k* clasificaciones.

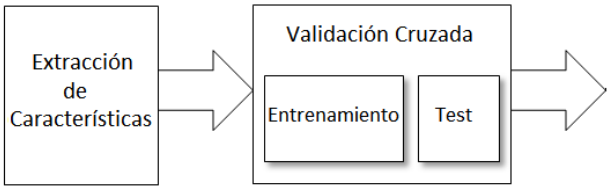

Figura 3.5 Etapas en el algoritmo de clasificación

3.3. Simulador de Conducción

Las simulaciones se han llevado a cabo en el simulador de conducción mostrado en la Figura 3.6, creado por el IoTLab de la Facultad de Informática, de la Universidad de Reutlingen, en Alemania. Y ha sido desarrollado con el software OpenDS, un simulador de conducción de código abierto, escrito en Java y desarrollado originalmente por el Centro Alemán de Investigación en Inteligencia Artificial (DFKI GmbH).

OpenDS consta de tres componentes principales que son el editor de tareas de conducción, el simulador y el analizador de unidades.

Con el editor gráfico de tareas de conducción, el usuario puede cargar un modelo de mapa vacío y colocar más elementos, como señales de tráfico, semáforos y vehículos. Además, se pueden especificar propiedades y eventos del automóvil,

que se activarán en el simulador en tiempo de ejecución. Y el analizador de conducción permite visualizar los datos del automóvil grabados durante una conducción varias veces por segundo, como pueden ser la posición, la dirección, la velocidad y el estado de los pedales (www.opends.eu).

Figura 3.6 Simulador del IoTLab de la Universidad de Reutlingen

3.3.1. Arquitectura del Simulador

La arquitectura del simulador puede verse en la Figura 3.7, donde los elementos en color rojo representan los dispositivos de entrada y en color verde los de salida.

Los dispositivos de salida se componen de tres pantallas para visualizar el entorno, colocadas delante del asiento del conductor, de manera que se crea una proyección de 180 grados. Y el cuadro de mandos se coloca detrás del volante y proporciona información sobre la velocidad de conducción y el nivel de combustible. Para facilitar una simulación más realista y permitir una mayor inmersión en la simulación, el sistema consta de cinco altavoces que rodean el asiento. Además, hay una pantalla táctil en el lado derecho del volante que representa la consola central del coche.

En cuanto a los dispositivos de entrada lo forman el volante, la palanca de cambios, y los pedales de aceleración, freno y embrague. Asimismo, hay un sensor de oído para monitorizar la frecuencia cardíaca y un electroencefalograma (EEG) para medir la actividad cerebral del conductor.

Además de las interfaces de entrada y salida para el conductor, el simulador consta de tres ordenadores que son responsables de la simulación del coche y el entorno, la recogida de datos del vehículo y del conductor, y la presentación de las aplicaciones desarrolladas para el conductor en la pantalla táctil.

3.3.2. Datos Generados

Los datos proporcionados por el simulador han sido muestreados cada segundo, de modo que se pueda trabajar con ellos con los algoritmos ya empleados con la base de datos real, y son los siguientes:

- *Timestamp* (Marca de tiempo)
- *Vehicle_Speed* (Velocidad del vehículo)
- *Engine_Speed* (Velocidad del motor)
- *GPS_Heading* (Rumbo GPS)
- *GPS_Longitude* (Longitud GPS)
- *GPS_Latitude* (Latitud GPS)
- *Engine_Acceleration* (Aceleración del motor)
- *Engine_Milage* (Kilometraje)
- *Engine_Engine_RPM* (Revoluciones por minuto del motor)
- *Gearbox_Position_Manual* (Posición manual de la caja de cambios)
- *Engine_Engine_Status* (Estado del motor)
- *Engine_Fuel_Consumption* (Consumo de combustible del motor)
- *Engine_Opt_Fuel_Consumption* (Consumo de combustible óptimo)
- *Engine_Petrol_Level* (Nivel de gasolina del motor)
- *Steering_Angle* (Ángulo de dirección)

Además, de los datos que se pueden obtener del encefalograma y del sensor de frecuencia cardíaca.

Como se verá en el siguiente capítulo, la única variable utilizada va a ser la velocidad del vehículo, a partir de la cual se determinará la distancia recorrida, así como los valores de aceleración en cada segundo.

Una vez visto las variables de datos con las que se ha trabajado, se presenta el algoritmo propuesto por Peña y Prieto para la detección de observaciones atípicas en bases de datos multivariantes, dado que éste ha sido el algoritmo utilizado en todas las pruebas junto con OCSVM, por considerarlo uno de los más eficientes.

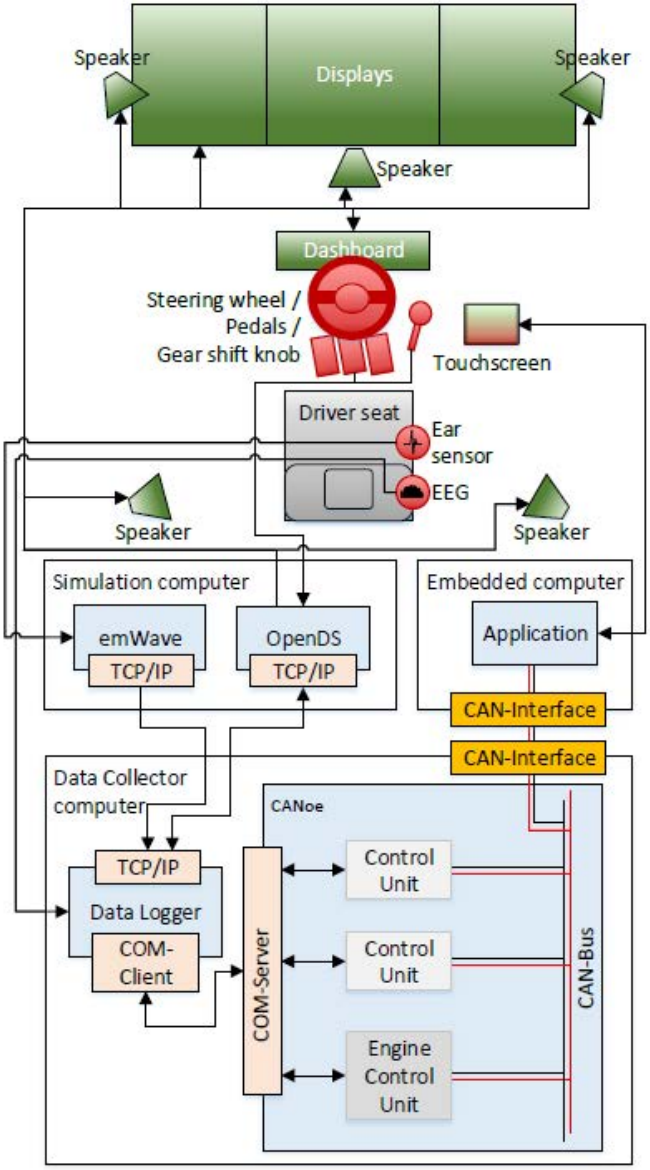

Figura 3.7 Arquitectura del simulador de conducción (Yay, 2.016)

3.4. Algoritmo de Peña y Prieto

En la detección de valores atípicos multivariantes no existe una solución universal y la elección del mejor método dependerá de diferentes parámetros, tales como el número de dimensiones que forma la estructura de los datos, o la clase o tipo de dichos datos. Es por tanto, la combinación de diferentes tipos de algoritmos, la mejor solución para optimizar los resultados en la detección de atípicos. En este caso, un ejemplo es el algoritmo desarrollado por Peña y Prieto (Peña, 2.001), donde se combinan las técnicas de búsqueda de proyecciones y las estadísticas con estimaciones robustas.

El método ideado por Peña y Prieto es un algoritmo iterativo, en el cual las observaciones sospechosas de ser atípicas son eliminadas de los datos originales. Con un conjunto de datos de p variables, los datos son proyectados en $2p$ direcciones, p direcciones de máxima curtosis y p direcciones de mínima curtosis.

Dado que cualquier observación atípica multivariante debe aparecer como atípica en al menos una dirección de proyección, la idea de usar direcciones que maximicen y minimicen el coeficiente de curtosis de las observaciones proyectadas, asegura encontrar estos atípicos, ya que por un lado, en variables univariantes el coeficiente de curtosis se incrementa por la presencia de valores atípicos, y por otro lado un grupo grande de atípicos puede causar bimodalidad y baja curtosis (Peña, 2.002).

Cada vez que una observación es eliminada, se calculan los estimadores robustos del resto de la muestra de datos. Estos estimadores son el vector de medias y la matriz de covarianzas, con los cuales se calcula la distancia de Mahalanobis de la observación con el centro de los datos y se considerará entonces como valor atípico, aquel que sea mayor a un cierto umbral. Bajo la hipótesis de normalidad multivariante, el cuadrado de la distancia de Mahalanobis sigue una distribución chi-cuadrado (χ^2) con p grados de libertad, de modo que un umbral razonable estaría por encima del cuantil 0,975.

Se ha observado experimentalmente que incluso cuando los datos no siguen estrictamente una distribución normal, la distribución χ^2 sigue siendo una buena aproximación.

3.4.1. Descripción del Algoritmo

1. Los datos de entrada $(x_1,\ldots,x_n)$, son transformados en una muestra estandarizada $(y_1,\ldots,y_n)$, de media cero y matriz de covarianzas la matriz identidad. Siendo $\bar{x}$ el vector de medias de la matriz de entrada y S su matriz de covarianzas, la muestra estandarizada Y se calcula como:

$$y_i = S^{-1/2}\left(x_i - \bar{x}\right)$$

2. Llamada a la función *d_kurtosis.R* para el cálculo de las p direcciones d_j, de norma unidad, que maximizan la curtosis de los datos proyectados sobre dichas direcciones.

   Cálculo de la proyección de la observación $y_i^{(j)}$ sobre la dirección d_j:

$$z_i^{(j)} = d_j^T \cdot y_i^{(j)}$$

3. La muestra Y se proyecta sobre un espacio de dimensión una unidad menos que los y_i, que será la muestra en la siguiente iteración.

   Se crea un vector $e_1 = (1,0,0,\ldots,0)$ de dimensión (p-j+1).

   a. Si $d_j = e_1$ se toma la nueva muestra como $y_i^{(j)} = \begin{pmatrix} z_i^{(j)} \\ y_i^{(j+1)} \end{pmatrix}$

   b. Si $d_j \neq e_1$ se definen el vector y la matriz:

$$v_j = d_j - e_1, \quad Q_j = I - \frac{v_j \cdot v_j^T}{v_j^T \cdot d_j}$$

   y para cada $y_i^{(j)}$ se obtiene $u_i^{(j)} = Q_j y_i^{(j)} = \begin{pmatrix} z_i^{(j)} \\ y_i^{(j+1)} \end{pmatrix}$

   c. Si $j = p \rightarrow z_i^{(p)} = y_i^p$

4. Se repite para las p direcciones d_j, que minimizan la curtosis, hasta obtener $2p$ direcciones.

5. Se eliminan todas las muestras sospechosas de ser valores atípicos, que serán aquellas que cumplan para algún *j*:

$$\frac{\left|z_i^{(j)} - med(z^{(j)})\right|}{Meda(z^{(j)})} > \beta_p$$

Siendo $med(z^{(j)})$ la mediana de los $z_i^{(j)}$, $Meda(z^{(j)})$ la mediana de $\left|z_i^{(j)} - med(z^{(j)})\right|$ y β_p elegido en función del número de variables.

6. Una vez eliminadas las observaciones sospechosas se calcula el vector de medias $\bar{x}_R$ y la matriz de covarianzas S_R robustos, es decir, para las observaciones x_i no eliminadas.

7. Se consideran atípicas todas aquellas observaciones cuya distancia de Mahalanobis entre la muestra y el vector de medias robusto cumplan:

$$d_R^2(x, \bar{x}_R) = (x - \bar{x}_R)^T S_R^{-1} (x - \bar{x}_R) > p + 3\sqrt{2p}$$

Los valores β_p, mostrados en la Tabla 3.3, han sido obtenidos mediante simulación con muestras sin datos atípicos, y para cualquier otro número de variables son calculados por interpolación lineal de los logaritmos de *p* y β_p.

Tabla 3.3 Valores de β_p en función del número de variables (Peña y Prieto, 2.001),

p	β_p
5	4,1
10	6,9
20	10,8

3.4.2. Función d_kurtosis.R

Para el cálculo de las direcciones d_j que maximizan o minimizan la curtosis de los datos proyectados sobre dicha dirección, se llama a la función *d_kurtosis.R*, que recibe como entrada la matriz Y de datos estandarizados y un índice para indicar si maximizamos o minimizamos.

El algoritmo es el siguiente:

1. Inicialización de $d_j = (1,0,...,0)$
2. Sea M la matriz cuadrada definida como: $M = \sum_{i=1}^{m} \left(d_j^T y_i^{(j)}\right)^2 y_i^{(j)} \left(y_i^{(j)}\right)^T$
3. Hacer $d_j^{ant} = d_j$, siendo d_j el vector propio de norma uno de M correspondiente con el mayor autovalor, en el caso de calcular la dirección de máxima curtosis, y el vector propio del menor autovalor si se quiere calcular la dirección de mínima curtosis.
4. Repetir los pasos (2) y (3) hasta que $\left\| d_j - d_j^{ant} \right\| < \varepsilon$, tomando $\varepsilon = 0{,}01$

Un algoritmo alternativo para encontrar los máximos y mínimos locales, se muestra a continuación y consistiría en calcular el gradiente de la curtosis de la matriz de entrada en, por ejemplo, diez direcciones generadas aleatoriamente, hasta encontrar la dirección que maximice o minimice el coeficiente de curtosis. Los resultados mediante este algoritmo son muy parecidos, aunque el tiempo de cálculo aumenta considerablemente con respecto al anterior.

1. Generación de 10 direcciones $d_1,\ldots,d_{10}$ aleatorias de norma uno

2. Inicializo dirección $d^{ant} = 0$ y coeficientes $\varepsilon = 0{,}01$ y $\alpha = 0{,}1$

3. Mientras $\left\| d_j - d_j^{ant} \right\| > \varepsilon$

   a. Cálculo del gradiente: $G = \sum_{i=1}^{n} 4 \cdot \left(d_j^T \cdot y_i^{(j)} \right)^3 \cdot y_i^{(j)}$

      a.1. Si $\|G\| > 1 \rightarrow \Delta = G / \|G\|$, si no $\Delta = G$

   b. $d_j^{ant} = d_j$

$$d_j = \frac{d_j^{ant} + (\alpha \cdot \Delta)}{\left\| d_j^{ant} + (\alpha \cdot \Delta) \right\|}$$

   c. $j = j + 1$

4. Cálculo del coeficiente de curtosis para las 10 direcciones d_j:

$$kur = \sum_{i=1}^{n} \left(d_j^T \cdot y_i^{(j)} \right)^4$$

5. Selección de la dirección cuyo coeficiente de curtosis es máximo o mínimo.

4. PRUEBAS Y RESULTADOS

EN este capítulo se detallan las distintas pruebas llevadas a cabo en este trabajo. Los resultados obtenidos se han logrado por medio tanto de datos reales, capturados con la aplicación *SmartDriver* y usada por varios conductores, así como por datos obtenidos bajo un entorno de simulación. A partir de esta información se va a generar una base de datos multidimensional con la que se pretende cuantificar el comportamiento del conductor, así como la detección de situaciones anómalas de tráfico.

Como se comentó en el capítulo anterior, las simulaciones se han llevado a cabo en el grupo de investigación IoTLab *(Internet of Things Laboratory)* de la Facultad de Informática de la Universidad de Reutlingen, en Alemania.

4.1. Resultados Experimentales en Escenarios Reales

Los resultados experimentales han sido obtenidos gracias a la adquisición de datos reales generados por diferentes conductores, y en distintos tipos de vía. Se han realizado diversos test, los cuales van a validar y corroborar las ideas presentadas en el capítulo anterior, y que se detallan a continuación.

En las dos primeras pruebas, los datos han sido recogidos en un tramo de la autovía M40 de Madrid, entre los kilómetros 21 y 27, y cuyo trayecto se muestra en la Figura 4.1.

4.1.1. Prueba 1: Cálculo de Atípicos Multivariantes

Esta primera prueba consiste, básicamente, en una comparación entre las diferentes técnicas de detección de atípicos multivariantes. Para ello, los datos han sido recogidos por un único conductor, durante quince días, y en ambos sentidos del tramo de autovía indicado.

Los datos capturados van a ser representados como una variable multivariante, formada a su vez por seis variables o componentes individuales, donde cada variable específica es procesada y relacionada con las otras componentes. Para crear esta variable multivariante se han considerado, por un lado las observaciones muestreadas cada segundo, y además las muestras generadas en intervalos de 30 segundos, dando lugar a un total de 3.808 observaciones durante los quince trayectos realizados.

Estas componentes o variables univariantes se muestran en la Tabla 4.1, y son la velocidad media por intervalo y la velocidad instantánea al final del intervalo, ambas medidas en m/s, la aceleración instantánea, medida en m/s^2, y las restantes variables son el PKE, RR y pNN50, ya explicadas en el capítulo anterior. En la Tabla 4.1 se pueden ver los estadísticos correspondientes a cada una de ellas, así como el número de atípicos univariantes encontrados.

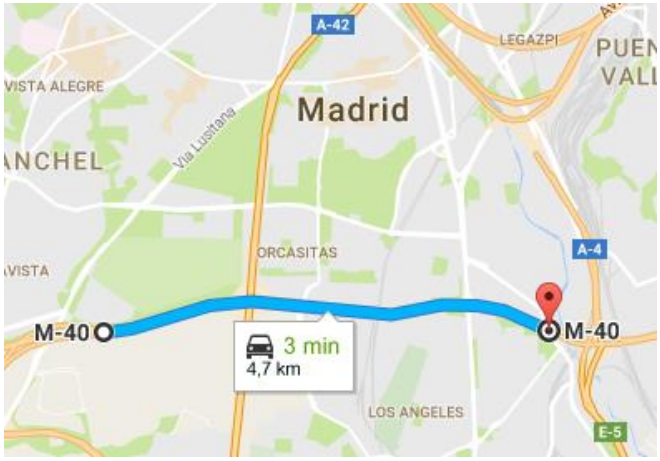

Figura 4.1 Tramo de conducción entre el km 21 y 27 de la autovía M40

Tabla 4.1 Estadísticos de las variables univariantes

Variable	Media	Desviación Típica	Mediana	Rango Intercuartílico	Atípicos
Velocidad Media	20,4373	7,93721	23,4118	10,4049	0
PKE	0,28136	0,20252	0,24608	0,20098	161
Velocidad	20,3893	8,14859	23,4082	10,2662	66
Aceleración	-0,00337	0,68909	0,00995	0,33969	286
RR	0,90758	0,11418	0,89200	0,15225	42
pNN50	25,5978	15,1336	23,3333	23,3333	2

El criterio seguido para obtener estos valores atípicos univariantes ha sido considerar como tales aquellos que distan más de 1,5 veces el rango intercuartil respecto de su primer y tercer cuartil.

Con el fin de detectar los atípicos multivariantes que aparecen en ese segmento de vía, han sido empleados varios métodos. En primer lugar se ha implementado el algoritmo de Peña y Prieto, así como el uso de un método robusto MCD. Con estas técnicas y bajo la hipótesis de normalidad multivariante, el cuadrado de la distancia de Mahalanobis se distribuye como una función chi-cuadrado, χ^2, con p grados de libertad, siendo p el número de variables. En este caso se van a considerar como observaciones atípicas todas aquellas cuya distancia al centro de los datos sea muy elevada El umbral usado en ambos métodos para determinar si una observación es considerada como atípica, ha sido una distancia estadística de 14,449, correspondiente al cuantil 0,975. Aunque algunos autores sugieren un umbral más conservativo como puede ser el cuantil 0,999 (Hair et al., 1999).

También se han implementado los algoritmos LOF y *clustering k-means*, obteniéndose resultados bastante similares, como se muestra en la Tabla 4.2, en la cual se pueden ver las 10 observaciones más alejadas para cada tipo de algoritmo utilizado, con sus respectivas distancias de Mahalanobis, y donde se aprecia claramente la existencia de cinco observaciones muy alejadas en relación al resto de los datos, tal y como se puede ver en la Figura 4.2. Además se tienen más de diez valores atípicos con distancias de Mahalanobis por encima de 500.

Llegado a este punto, es importante tener en cuenta que es el investigador quien va a decidir si una observación debe considerarse como atípica o no, con independencia de la distancia a la que se encuentre.

Tabla 4.2 Los 10 atípicos más alejados obtenidos con diferentes algoritmos

Peña y Prieto		MCD		LOF (k=7)	Clustering k-means (k=7)
Observación	**Distancia de Mahalanobis**	**Observación**	**Distancia de Mahalanobis**	**Observación**	**Observación**
3.280	7.229,058	3.251	4.590,819	3.252	3.251
3.281	5.943,931	3.252	4.022,075	3.251	3.252
551	2.525,677	522	1.618,883	1.993	522
1.257	1.007,746	1.228	642,338	2.185	1.228
2.562	809,454	2.533	517,518	522	2.533
1.459	791,944	1.430	505,688	523	1.430
711	582,307	3.607	373,883	3.591	682
3.636	552,928	682	371,494	626	1.993
2.022	513,706	1.993	341,520	2.533	3.607
3.635	504,233	3.606	334,579	3.602	3.608

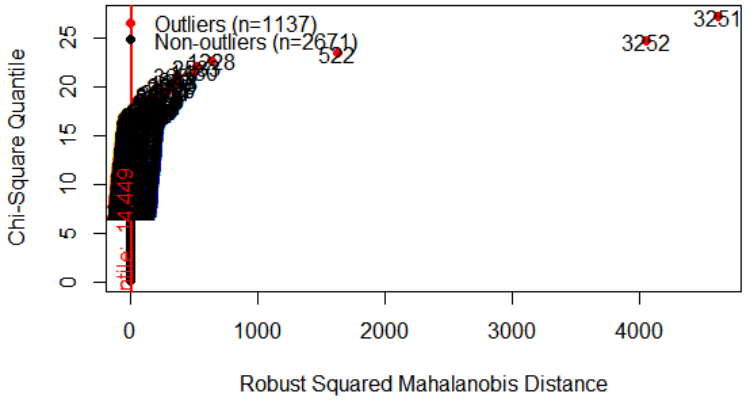

Figura 4.2 Atípicos detectados mediante MCD

Otro de los métodos empleados como detector de atípicos ha sido *One-class SVM*. En la Tabla 4.3 se muestran el número total de atípicos multivariantes encontrados para cada método, obteniéndose un resultado cercano en el caso de usar el algoritmo de Peña y Prieto y el método robusto MCD, siendo 1.282 y 1.137 atípicos detectados respectivamente, y 762 atípicos con la SVM de una clase, del total de las 3.808 observaciones. Claramente el uso de OCSVM proporciona un resultado más restrictivo, motivo por el cual es una de las técnicas utilizadas.

A pesar de que en los dos primeros métodos se asume una distribución normal de la variable multivariante y este requisito no se cumple, aun así los resultados son muy similares con los obtenidos mediante técnicas de minería de atípicos como los algoritmos LOF y *clustering k-means*, al menos para las observaciones más alejadas, y tomando por ejemplo siete grupos.

Por otra parte, es interesante puntualizar que los atípicos multivariantes no aparecen en el análisis univariante, como se puede observar en la Tabla 4.1. Esto pone de manifiesto que cada atípico multivariante no es único en cada variable individual o univariante, sino que es único en la combinación de variables (Hair et al., 1999).

Observando la Tabla 4.4, se ve que las variables correlacionadas son la velocidad y la velocidad media. También aparece una correlación alta entre las variables RR y pNN50, como era de esperar. Siendo las correlaciones del resto de componentes muy bajas o prácticamente nulas.

En el caso de ser eliminadas estas variables correlacionadas se ha comprobado que los resultados seguirían manteniendo las mismas cinco observaciones atípicas más alejadas, aunque el número total de atípicos sería menor.

Tabla 4.3 Número de atípicos multivariantes encontrados por cada método

Técnica	Atípicos/Observaciones	% Atípicos
Algoritmo de Peña y Prieto	1.282/3.808	33,66%
Algoritmo MCD	1.137/3.808	29,85%
SVM de una clase	762/3.808	20%

Tabla 4.4 Matriz de correlaciones. Prueba 1

	Vel. Media	PKE	Velocidad	Aceleración	RR	pNN50
Vel. Media	1,00000	-0,31251	0,94982	-0,06084	-0,36778	-0,35041
PKE	-0,31251	1,00000	-0,14031	0,14522	0,04317	0,08515
Velocidad	0,94982	-0,14031	1,00000	0,04211	-0,35171	-0,34728
Aceleración	-0,06084	0,14522	0,04211	1,00000	0,02244	0,00522
RR	-0,36778	0,04317	-0,35171	0,02244	1,00000	0,63824
pNN50	-0,35041	0,08515	-0,34728	0,00522	0,63824	1,00000

Dado que las diferentes observaciones atípicas encontradas deben corresponderse con condiciones singulares acontecidas durante la conducción, se va a considerar que, al ser un tramo de autovía, estos atípicos podrían representar la aparición de situaciones de congestión de tráfico o retenciones.

En la prueba que se detalla a continuación se va a contrastar la relación entre tales valores atípicos y la posibilidad de que se hayan producido atascos.

4.1.2. Prueba 2: Clasificación de Trayectos con Atasco

En esta segunda prueba se va a evaluar la probabilidad de que los valores atípicos detectados correspondan, o no, con una situación de retención. Para ello se va a hacer uso de varios clasificadores. En este caso, los datos usados corresponden al mismo tramo de autovía, pero son recogidos por dos conductores diferentes, aunque no se han tenido en cuenta los datos de frecuencia cardíaca.

Para calcular el número de atípicos por día se han usado cuatro de las seis variables empleadas en la prueba anterior. Estas variables son las velocidades media e instantánea, la aceleración PKE y la aceleración instantánea.

En total se dispone de una muestra de 32 días, 19 de ellos correspondientes al conductor 1 y los restantes 13 días al conductor 2, y donde 5 de los días han sido etiquetados como atasco por los propios conductores.

En primer lugar, el método utilizado para detectar los valores atípicos va a ser

el algoritmo de Peña y Prieto, tomando como umbral de corte para la distancia de Mahalanobis, el cuantil 0,999. Dicho cuantil se corresponde, para los cuatro grados de libertad o variables individuales, con una distancia estadística de 18,4662.

Las variables que se han tenido en cuenta para entrenar los clasificadores han sido:

- *número total de atípicos por día*
- *distancia máxima de Mahalanobis*
- *velocidad mínima*
- *distancia media de Mahalanobis*
- *velocidad media*

Y se muestran en la Tabla 4.5, donde aparecen los datos para cada conductor y en cada uno de los días evaluados.

Observando la matriz de correlaciones entre estas cinco variables en la Tabla 4.6, en primer lugar, se puede ver que existe una alta correlación entre el número de atípicos que suceden en un día con el hecho de que ocurra un atasco.

Asimismo, como cabría esperar, hay una alta correlación entre las situaciones de atasco con las variables de velocidad mínima y media, mientras que la correlación entre situaciones de atasco con la distancia media de Mahalanobis es totalmente inexistente, y sí tiene una baja correlación con la distancia máxima de Mahalanobis.

Por otro lado, el número de atípicos por día presenta una alta correlación con las velocidades mínima y media. También tiene una alta correlación con la máxima distancia de Mahalanobis, aunque esa correlación es nula en cuanto a su distancia media.

La distancia media de Mahalanobis solo tiene una baja correlación con la distancia máxima de Mahalanobis y está totalmente incorrelada con el resto de variables.

Y el valor de mayor correlación aparece, como es normal, entre la velocidad mínima y la velocidad media.

Tabla 4.5 Días con atasco por cada conductor con datos de entrada al clasificador

Día	Conductor	Atípicos	Máx. DM	Media DM	Vel. Mín.	Vel. Media	Atasco
28-07-2016	1	2	81,07729	66,55632	20,74825	21,44664	0
13-09-2016	1	1	25,53228	25,53228	21,70138	21,70138	0
19-09-2016	1	3	84,43930	42,64661	20,44467	21,82940	0
23-09-2016	1	15	52,19958	26,89702	19,24631	23,12502	0
24-09-2016	1	5	69,60313	34,37197	21,82484	23,57318	0
27-09-2016	1	4	733,1730	202,7303	19,23984	20,32243	0
30-09-2016	1	49	1023,753	81,84298	3,190017	12,54775	1
01-10-2016	1	35	6425,263	369,4148	24,32699	27,37210	0
04-10-2016	1	11	216,0537	53,58827	17,99887	20,55648	0
05-10-2016	1	6	161,4753	49,31944	20,96285	24,27076	0
22-10-2016	1	38	826,8321	53,54304	15,78341	20,72422	0
05-11-2016	1	37	1289,638	68,98017	16,66671	21,34352	0
07-11-2016	1	42	433,1340	38,05313	20,41196	27,37902	0
08-11-2016	1	13	260,2814	49,86332	19,74296	22,19054	0
30-11-2016	1	160	487,6518	61,05510	3,841789	12,08670	1
13-12-2016	1	55	623,7417	70,25145	9,515182	21,81178	0
10-01-2017	1	11	413,5456	61,87721	20,93726	24,13760	0
11-02-2017	1	4	950,6243	257,0559	21,44350	24,75372	0
28-02-2017	1	31	472,1541	54,86111	17,88974	26,09242	0
18-10-2016	2	20	63,9782	32,07765	18,24054	23,24320	0
19-10-2016	2	4	75,02392	49,60632	23,73005	24,85006	0
20-10-2016	2	12	207,6386	44,34776	20,53107	25,48546	0
28-10-2016	2	13	160,5586	42,20218	19,67294	22,60530	0
08-11-2016	2	258	1350,503	68,71632	1,210041	8,695109	1
23-11-2016	2	8	247,5452	63,93788	18,45000	21,81625	0
13-12-2016	2	7	870,5738	147,0208	20,33000	23,48714	0
01-02-2017	2	5	570,5017	141,3194	24,26000	25,66000	0
15-02-2017	2	3	121,0327	55,05563	20,72000	21,01000	0
21-02-2017	2	11	1383,859	163,7110	16,91000	19,49455	0
23-02-2017	2	251	141,7401	48,19856	5,070000	10,10375	1
27-02-2017	2	1	19,84990	19,84990	26,33000	26,33000	0
02-03-2017	2	530	12251,53	126,1290	0,000000	7,595019	1

Los dos tipos de clasificadores empleados han sido el modelo de regresión logística o Logit y una SVM con kernel lineal.

Dado que el tamaño de los datos no es lo suficientemente elevado para entrenar los clasificadores, se ha usado validación cruzada de *k* iteraciones, de manera que todos los datos puedan usarse tanto para test como para entrenamiento. Se ha tomado un valor de $k = 4$, con lo que se tiene un 25% de los datos para test y un 75% para entrenamiento, como queda reflejado en la Tabla 4.7.

Los resultados obtenidos por ambos clasificadores han sido bastante similares, con una tasa de acierto del 93,75% para el clasificador Logit, mientras que la SVM ha conseguido un 100% de tasa de acierto, con unos coeficientes kappa, o índice de exactitud, de 0,763 y 1 respectivamente.

Y en las Tablas 4.8 y 4.9 se muestran, respectivamente, las matrices de confusión, esto es, las tablas de aciertos y errores, para ambos clasificadores, donde se puede apreciar si los valores predichos coinciden con los valores reales. Para el modelo Logit se obtiene un falso positivo y un falso negativo.

Tabla 4.6 Matriz de correlaciones. Prueba 2

	Atípicos	**Máx. DM**	**DM Media**	**Vel. Mín.**	**Vel. Media**	**Atasco**
Atípicos	1,00000	0,73235	0,03528	- 0,79409	- 0,79884	0,79811
Máx. DM	0,73235	1,00000	0,47968	- 0,38347	- 0,37648	0,38193
DM Media	0,03528	0,47968	1,00000	0,09223	0,08873	- 0,03650
Vel. Mín.	- 0,79409	- 0,38347	0,09223	1,00000	0,92356	- 0,90350
Vel. Media	- 0,79884	- 0,37648	0,08873	0,92356	1,00000	- 0,91128
Atasco	0,79811	0,38193	- 0,03650	- 0,90350	- 0,91128	1,00000

Tabla 4.7 Validación cruzada de *k* iteraciones con $k = 4$

Test	Entrenamiento		
Entrenamiento	Test	Entrenamiento	
Entrenamiento		Test	Entrenamiento
Entrenamiento			Test

Como se puede advertir, los resultados son muy buenos, pero no excesivamente significativos, ya que el número de datos con los que se ha alimentado a los clasificadores para su entrenamiento no ha sido lo suficiente, de lo que cabría desear.

Tabla 4.8 Matriz de confusión para el modelo Logit

Real / Predicción	Sin atasco	Atasco	Actual	Sensibilidad
Sin atasco	26	1	27	96,3%
Atasco	1	4	5	80%
Predicho	27	5	32	88,15% Sensibilidad media
Precisión	96,3%	80%	88,15% Precisión media	93,75% Exactitud media

Tabla 4.9 Matriz de confusión para la SVM

Real / Predicción	Sin atasco	Atasco	Actual	Sensibilidad
Sin atasco	27	0	27	100%
Atasco	0	5	5	100%
Predicho	27	5	32	100% Sensibilidad media
Precisión	100%	100%	100% Precisión media	100% Exactitud media

La tasa media de acierto o exactitud media es la proporción del número total de predicciones que son correctas, y se calcula como:

$$Exactitud = \frac{TP+TN}{TP+TN+FP+FN} \tag{4-1}$$

Donde TP son los positivos verdaderos o *true positives* y son el número de casos en que el modelo predice la clase de interés (o clase positiva) y acierta.

TN representan los negativos verdaderos o *true negatives* y son el número de

casos en que el modelo predice las clases distintas de la de interés (o clases negativas) y acierta. FP son los falsos positivos o *false positives* y son los casos en que el modelo predice la clase positiva erróneamente. Y por último, FN son los falsos negativos o *false negatives* y son los casos en que el modelo predice las clases negativas erróneamente.

La sensibilidad o *recall* indica la fracción de instancias relevantes que han sido recuperadas. Y la precisión es la fracción de instancias recuperadas que son relevantes.

Para la clase positiva la sensibilidad y la precisión vendrán dadas por:

$$Sensibilidad = \frac{TP}{TP+FN} \tag{4-2}$$

$$Precisión = \frac{TP}{TP+FP} \tag{4-3}$$

Mientras que para la clase negativa la sensibilidad y la precisión serán:

$$Sensibilidad = \frac{TN}{TN+FP} \tag{4-4}$$

$$Precisión = \frac{TN}{TN+FN} \tag{4-5}$$

4.1.2.1. Clasificación de atascos en función sólo de valores atípicos, obtenidos con el algoritmo de Peña y Prieto

Dado que la velocidad es una variable íntegramente relacionada con la probabilidad de que haya o no un atasco, resulta mucho más interesante realizar estos test usando variables que sean calculadas solo en base a las observaciones atípicas.

En este caso las variables elegidas han sido:

- *número total de atípicos por día*
- *densidad máxima de atípicos en un tramo de autovía*
- *número de ráfagas de atípicos en ese tramo superior a 10*

El tramo de autovía considerado tiene una longitud de aproximadamente 200 metros. Y como número de ráfagas de atípicos se entiende el número de veces que aparecen más de 10 atípicos seguidos en esos 200 metros.

En este caso, la matriz de correlación, en la Tabla 4.10, muestra efectivamente cómo la presencia de un atasco está altamente correlacionada con la máxima densidad de atípicos en un tramo de autovía, así como con el número de ráfagas superiores a 10 atípicos.

Nuevamente se han utilizado los clasificadores Logit y SVM, obteniéndose idénticos resultados por ambos, con una tasa media de acierto del 96,88%, y un coeficiente kappa de 0,871. La matriz de confusión se muestra a continuación en la Tabla 4.11, donde se tiene únicamente un falso positivo.

Tabla 4.10 Matriz de correlaciones. Prueba 2a

	Atípicos	**Densidad Máx.**	**Nº de Ráfagas**	**Atasco**
Atípicos	1,0000000	0,9711965	0,9612990	0,7981140
Densidad Máx.	0,9711965	1,0000000	0,9518459	0,8237436
Nº de Ráfagas	0,9612990	0,9518459	1,0000000	0,8713098
Atasco	0,7981140	0,8237436	0,8713098	1,0000000

Tabla 4.11 Matriz de confusión del modelo Logit y SVM. Prueba 2a

Real / Predicción	**Sin atasco**	**Atasco**	**Actual**	**Sensibilidad**
Sin atasco	27	1	28	96,43%
Atasco	0	4	4	100%
Predicho	27	5	32	98,21% Sensibilidad media
Precisión	100%	80%	90% Precisión media	96,88% Exactitud media

4.1.2.2. Clasificación de atascos en función sólo de valores atípicos, obtenidos con SVM de una clase

Este mismo test ha sido realizado calculando el número total de atípicos por día con una SVM de una clase. En este caso, la matriz de correlación, en la Tabla 4.12, muestra una mayor correlación entre la densidad máxima de atípicos y los atascos, aunque algo menor que en el test anterior.

Los resultados han empeorado ligeramente para el clasificador Logit, como se aprecia en la Tabla 4.13, con una tasa media de acierto del 93,75%, y un coeficiente kappa de 0,763. Para la SVM el resultado es el mismo que para el algoritmo de Peña y Prieto, Tabla 4.14, con una tasa media de acierto del 96,88% y coeficiente kappa de 0,871.

Otra vez se han obtenido muy buenos resultados, y aunque los datos de tráfico real recogidos corresponden a unos pocos días, cabe considerar que hay bastantes indicios que ponen de manifiesto que es un enfoque viable y que puede ser bastante efectivo.

Tabla 4.12 Matriz de correlaciones. Prueba 2b

	Atípicos	**Máx. Densidad**	**Nº de Ráfagas**	**Atasco**
Atípicos	1,0000000	0,9579154	0,9598384	0,7612141
Máx. Densidad	0,9579154	1,0000000	0,9783468	0,7171856
Nº de Ráfagas	0,9598384	0,9783468	1,0000000	0,6714705
Atasco	0,7612141	0,7171856	0,6714705	1,0000000

Tabla 4.13 Matriz de confusión del modelo Logit. Prueba 2b

Real / Predicción	**Sin atasco**	**Atasco**	**Actual**	**Sensibilidad**
Sin atasco	26	1	27	96,3%
Atasco	1	4	5	80%
Predicho	27	5	32	88,15% Sensibilidad media
Precisión	96,3%	80%	88,15% Precisión media	93,75% Exactitud media

Tabla 4.14 Matriz de confusión para SVM. Prueba 2b

Real / Predicción	Sin atasco	Atasco	Actual	Sensibilidad
Sin atasco	27	1	28	96,43%
Atasco	0	4	4	100%
Predicho	27	5	32	98,21% Sensibilidad media
Precisión	100%	80%	90% Precisión media	96,88% Exactitud media

4.1.3. Prueba 3: Detección de Rotondas y Pasos de Cebra en Leganés

Esta tercera prueba tiene como objetivo la identificación de diferentes elementos de la infraestructura vial urbana, por medio de la detección de atípicos de una variable individual o univariante.

Se han utilizado los datos recogidos en un tramo urbano de unos 2,5 km de longitud, como se muestra en la Figura 4.3. Dichos datos corresponden a un único conductor, durante 15 trayectos y recorridos en el mismo sentido de circulación.

En este tramo aparecen distintos elementos de infraestructura vial como son rotondas, pasos de cebra, señales de tráfico y semáforos, aunque los resultados se han centrado en varias clases particulares como son las rotondas y los pasos de cebra.

En este caso, la variable utilizada ha sido la aceleración del vehículo, calculada a partir de la velocidad, tal y como se indica en la ecuación (4-6).

La velocidad es muestreada cada segundo, de modo que se va a calcular la aceleración media mediante la diferencia de velocidades entre dos muestras consecutivas. Este valor de aceleración puede considerarse como la aceleración instantánea en cada observación, medida en m/s^2.

$$a = \frac{\Delta v}{\Delta t} = \frac{v_i - v_{i-1}}{t_i - t_{i-1}} \tag{4-6}$$

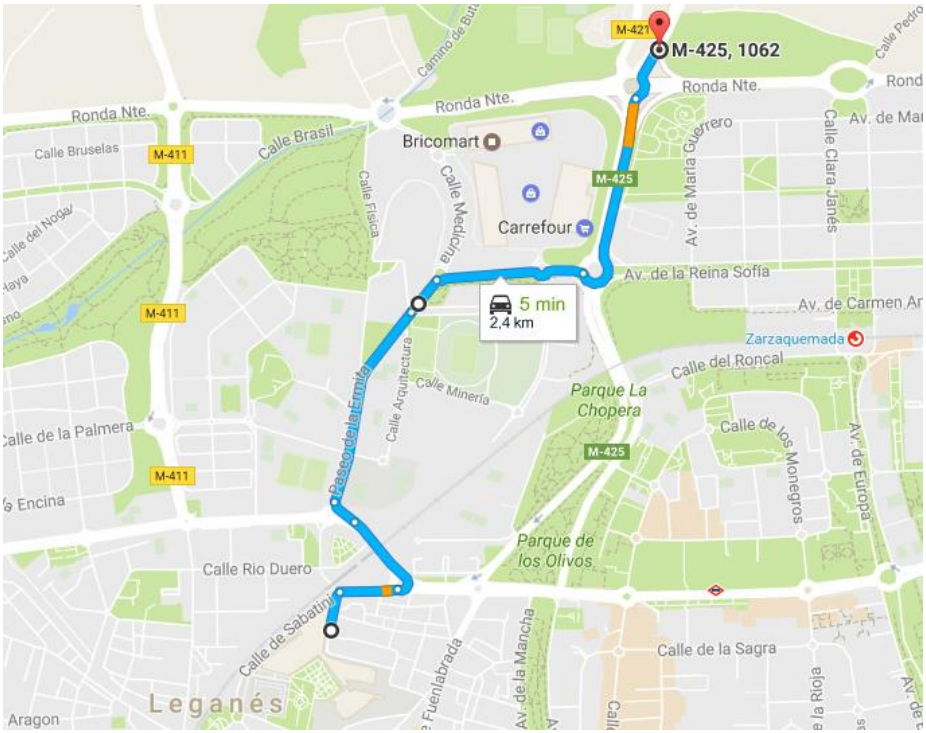

Figura 4.3 Tramo urbano de Leganés

Como ejemplo, la Tabla 4.15 muestra el resultado de la detección de atípicos para un día concreto en el tramo indicado. El recorrido consta de un total de 273 observaciones de las cuales se han encontrado 18 atípicos para la aceleración. La última columna de la tabla indica el número de la observación que es atípica.

En la Figura 4.4 se representan los valores de aceleración, donde se pueden apreciar claramente la presencia de los 18 picos correspondientes con valores atípicos.

Para calcular estos datos atípicos se ha considerado cada punto con las observaciones que aparecen en los 10 segundos precedentes y los 10 segundos posteriores, de modo que cada muestra va a constar de 21 observaciones.

En este caso, como criterio de atipicidad se va a calcular el *boxplot* de cada muestra, tomando de nuevo como margen dos veces el valor del rango intercuartil, con respecto al primer y tercer cuartil.

Tabla 4.15 Atípicos en la aceleración del día 10-01-2017

Marca de Tiempo	Latitud	Longitud	Velocidad	Aceleración	Atípico
2017-01-10 21:24:14+01:00	40,33455044	-3,76465642	0,00000000	-3,5187707	15
2017-01-10 21:24:15+01:00	40,33460932	-3,76459320	3,99099707	3,9909970	16
2017-01-10 21:25:06+01:00	40,33684969	-3,764858515	3,53432655	1,9888235	65
2017-01-10 21:25:10+01:00	40,33709265	-3,764787137	7,85306024	3,6293082	69
2017-01-10 21:25:22+01:00	40,33824145	-3,764337774	10,8394002	2,2800436	81
2017-01-10 21:25:25+01:00	40,33855584	-3,764232706	11,1690139	1,2059478	84
2017-01-10 21:25:27+01:00	40,33877438	-3,764160691	12,1093702	0,9403562	86
2017-01-10 21:25:41+01:00	40,34026915	-3,763600455	11,3718643	1,0312309	99
2017-01-10 21:26:05+01:00	40,34218731	-3,761678822	8,27513790	-2,6763477	121
2017-01-10 21:26:43+01:00	40,34250625	-3,758467995	1,10592150	1,1059215	160
2017-01-10 21:26:44+01:00	40,34247632	-3,758411906	0,00000000	-1,1059215	161
2017-01-10 21:26:45+01:00	40,34244998	-3,758372059	3,56356048	3,5635604	162
2017-01-10 21:27:03+01:00	40,34247797	-3,756960586	4,78179454	-1,1428757	180
2017-01-10 21:27:04+01:00	40,34247797	-3,756960586	0,00000000	-4,7817945	181
2017-01-10 21:27:27+01:00	40,34248097	-3,756906198	2,31053829	2,3105382	204
2017-01-10 21:27:28+01:00	40,34244823	-3,756838277	0,00000000	-2,3105383	205
2017-01-10 21:27:29+01:00	40,34238898	-3,756792208	6,13154077	6,1315407	206
2017-01-10 21:27:30+01:00	40,34238898	-3,756792208	7,91656160	1,7850208	207

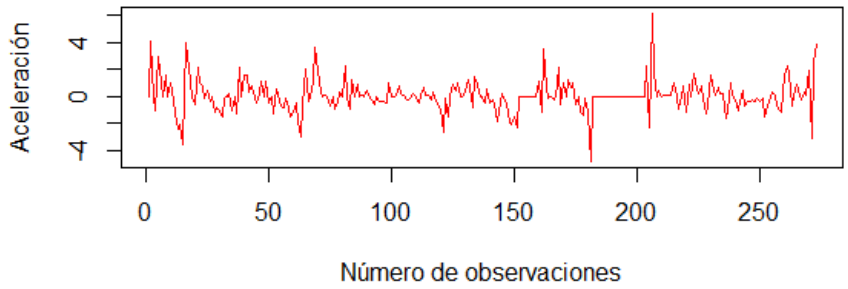

Figura 4.4 Valores de aceleración durante el trayecto indicado el 10-01-2017

A partir de la detección de las observaciones atípicas de la aceleración en cada uno de los quince días, y en combinación con el uso de clasificadores, se va a proceder a la identificación de las rotondas y pasos de cebra que se encuentran en ese tramo, como se explicará más adelante.

Y en la Tabla 4.16, se muestra el número de atípicos obtenidos para cada uno de los quince días.

La idea que se quiere subrayar es que, algunos de estos valores que son atípicos en cada trayecto, se deben corresponder a situaciones anómalas durante la conducción, como pueda ser la entrada a una rotonda o la parada en un paso de cebra.

En la Figura 4.5 se han representado las cinco rotondas y pasos de cebra existentes en ese tramo de vía, junto a sus respectivas coordenadas geográficas.

Tabla 4.16 Valores atípicos de la aceleración para cada día

Día	Fecha	Observaciones	Atípicos	% Atípicos
1	26-02-2016	365	10	2,74%
2	29-02-2016	287	7	2,44%
3	03-03-2016	371	22	5,93%
4	11-03-2016	298	17	5,70%
5	20-04-2016	326	15	4,60%
6	13-09-2016	321	18	5,60%
7	19-09-2016	317	17	5,36%
8	23-09-2016	329	19	5,77%
9	27-09-2016	329	14	4,25%
10	04-10-2016	308	13	4,22%
11	05-10-2016	292	15	5,13%
12	07-11-2016	331	25	7,55%
13	08-11-2016	294	23	7,82%
14	13-12-2016	318	19	5,97%
15	10-01-2017	273	18	6,59%

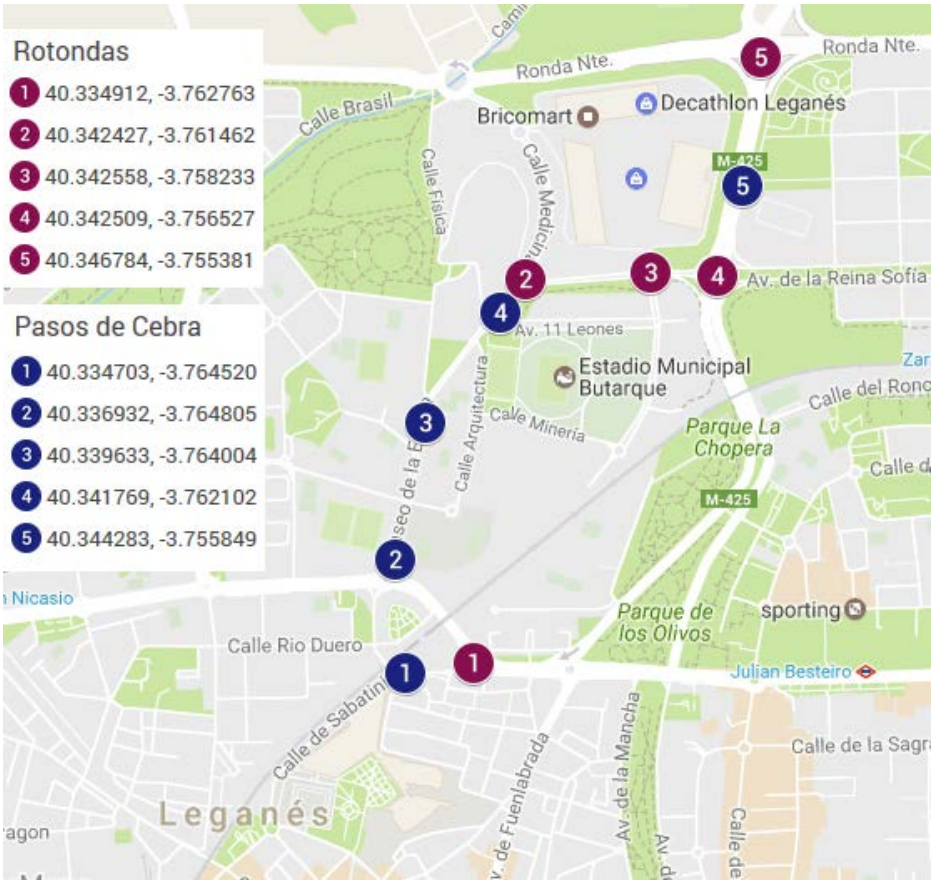

Figura 4.5 Rotondas y pasos de cebra en el tramo urbano de Leganés

Para obtener los atípicos correspondientes a cada una de las rotondas se van a considerar los puntos atípicos que disten menos de 50 metros respecto a las coordenadas de cada rotonda. Y para el caso de los pasos de cebra se ha considerado una distancia inferior a 16 metros, de manera que tomando unos 4 metros de longitud media por vehículo, se tendrían como mucho unos tres coches por delante.

Estas distancias han sido calculadas como se indica en la ecuación (4-7), utilizando la Ley Esférica del Coseno, donde Δs representa la distancia entre los puntos 1 y 2,

φ_1 y φ_2 son las latitudes expresadas en radianes para dichos puntos y $\Delta\delta$ es la diferencia de sus longitudes, también en radianes y *R* es el valor del radio medio de la Tierra.

Dado que esta fórmula calcula la distancia en línea recta entre ambos puntos y además asume que la Tierra es completamente redonda, cabe esperar una tasa de error, la cual se va a considerar despreciable en relación a las distancias calculadas.

$$\Delta s = \cos^{-1}(sen\varphi_1 \cdot sen\varphi_2 + \cos\varphi_1 \cdot \cos\varphi_2 \cdot \cos\Delta\delta)R \qquad (4\text{-}7)$$

Una vez obtenidos el número de datos atípicos para cada uno de los elementos urbanos, se van a etiquetar, junto con su velocidad media, y van a alimentar los clasificadores Logit y SVM con kernel lineal.

Hay que indicar que el número total de atípicos en cada infraestructura es inferior al de atípicos por día debido, entre otras causas, a que se han eliminado los atípicos cuyas coordenadas geográficas son iguales, ya que éstos se corresponden con situaciones en las que el vehículo está parado, pero el sensor continua enviando datos cada segundo, y por lo tanto se van a considerar todos esos puntos como un único valor atípico.

El archivo de datos que alimenta a los clasificadores se muestra en la Tabla 4.17, donde las clases cero y uno se corresponden con las rotondas y los pasos de cebra respectivamente, tal y como se puede ver en la Figura 4.6. De estos datos se evidencia que el número de atípicos es bastante más elevado en una rotonda que en un paso de cebra, mientras que el valor de la velocidad no es significativo.

Para entrenar los clasificadores, como el tamaño de los datos de entrada no es lo suficientemente elevado, se ha usado de nuevo validación cruzada de *k* iteraciones, de modo que todos los datos se usen tanto para test como para entrenamiento y así mejorar el rendimiento.

Se ha tomado un valor de $k = 5$, de manera que se tenga un 80% de los datos de entrada para entrenamiento y un 20% para test.

Las matrices de confusión se muestran en las Tablas 4.18 y 4.19, para la clasificación Logit y SVM respectivamente, cuyas tasas de acierto han sido del 100% y 90%, con coeficientes kappa de 1 y 0,8.

Una vez más, los resultados son muy buenos, pero la cantidad de datos parece insuficiente.

Tabla 4.17 Valores de entrada al clasificador. Prueba 3

Atípicos	Velocidad Media	Clase
17	4,346265	0
15	7,562202	0
19	4,048743	0
31	3,712298	0
26	2,003264	0
3	5,948125	1
10	4,700137	1
1	11,96490	1
4	3,970085	1
1	8,404808	1

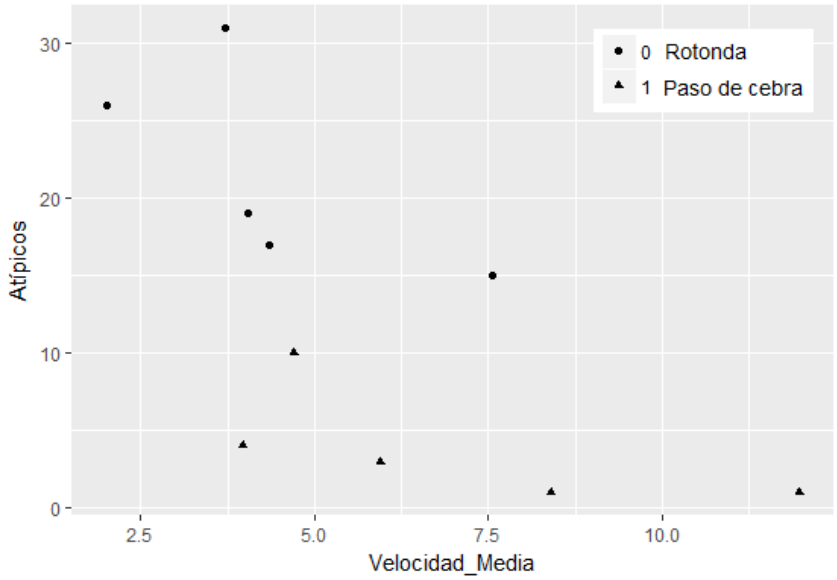

Figura 4.6 Clases de entrada al clasificador. Prueba 3

Tabla 4.18 Matriz de confusión para Logit: Rotonda-Paso de cebra

Real / Predicción	Rotonda	P.Cebra	Actual	Sensibilidad
Rotonda	5	0	5	100%
P.Cebra	0	5	5	100%
Predicho	5	5	10	100% Sensibilidad media
Precisión	100%	100%	100% Precisión media	100% Exactitud media

Tabla 4.19 Matriz de confusión para SVM: Rotonda-Paso de cebra

Real / Predicción	Rotonda	P.Cebra	Actual	Sensibilidad
Rotonda	4	0	4	100%
P.Cebra	1	5	6	83,33%
Predicho	5	5	10	91,67% Sensibilidad media
Precisión	80%	100%	90% Precisión media	90% Exactitud media

4.1.4. Prueba 4: Detección de Cruces y Rotondas en Stuttgart

En esta cuarta prueba se han usado los datos correspondientes a la base de datos del HCILab (*Human Computer Interaction Laboratory*) de la Universidad de Stuttgart (https://www.hcilab.org/research/hcilab-driving-dataset/).

Los datos se corresponden a conducciones en un mismo recorrido por parte de 10 conductores distintos, y forman un total de 14.048 observaciones para los diez recorridos realizados, en ese tramo urbano (Schneegass et al., 2013). En la Figura 4.7 se muestra el trayecto total, muestreado cada segundo, donde se aprecian varios cortes que se corresponden con conducción bajo túneles, donde la señal de GPS se ve interrumpida.

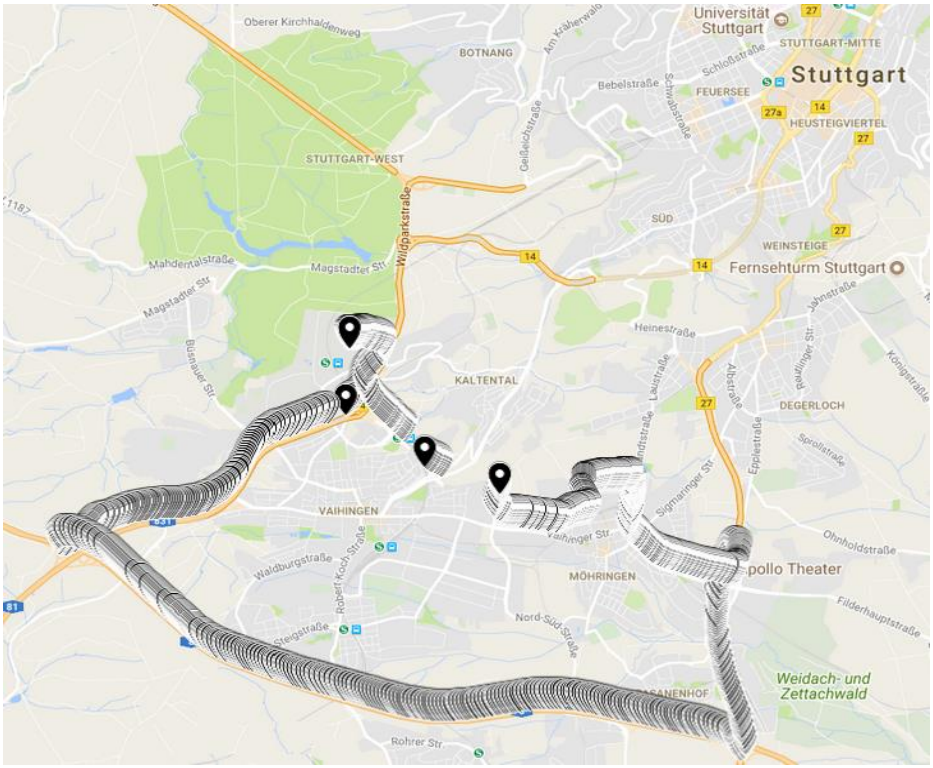

Figura 4.7 Recorrido del tramo de Stuttgart

Para calcular los valores atípicos de la aceleración generados en cada recorrido, al igual que se hizo en la prueba anterior, se toma cada punto con las observaciones que aparecen en los 10 segundos precedentes y los 10 segundos posteriores, de modo que cada muestra consta de 21 observaciones. Y como criterio de atipicidad se va a calcular el *boxplot* de cada muestra, tomando como margen dos veces el valor del rango intercuartil, con respecto al primer y tercer cuartil. El número de atípicos obtenidos para cada día se muestran a continuación en la Tabla 4.20.

En la Figura 4.8 se muestra en la parte superior el tramo urbano del trayecto considerado, para la detección de los elementos urbanos correspondientes a rotondas y cruces. Y en la parte inferior se señalan las coordenadas de cada elemento, marcadas en azul oscuro las dos rotondas y en azul más claro los ocho cruces, cuyas coordenadas se muestran además en la Tabla 4.21.

Para identificar el número de atípicos correspondiente a cada cruce se considera una distancia máxima de 14 metros respecto a éste, lo que equivaldría a unos 3 coches parados por delante. Y en el caso de las rotondas se ha considerado una distancia máxima de 60 metros desde el centro de la rotonda, donde cada una de ellas tiene un diámetro medio de unos 30 metros.

Tabla 4.20 Número de atípicos de la aceleración en cada recorrido

Conductor	Observaciones	Atípicos	% Atípicos
1	1515	36	2,37%
2	1431	20	1,39%
3	1457	38	2,60%
4	1501	26	1,73%
5	1321	41	3,10%
6	1382	30	2,17%
7	1428	30	2,10%
8	1254	15	1,19%
9	1379	32	2,32%
10	1380	27	1,95%

Tabla 4.21 Coordenadas GPS de los cruces y rotondas en el tramo de Stuttgart

Infraestructura Urbana	Latitud	Longitud
Cruce 1	48,726035	9,151684
Cruce 2	48,728722	9,150424
Cruce 3	48,732390	9,148709
Cruce 4	48,732263	9,145746
Cruce 5	48,731184	9,143122
Cruce 6	48,730667	9,142939
Cruce 7	48,729066	9,143971
Cruce 8	48,727829	9,139536
Rotonda 1	48,727097	9,150001
Rotonda 2	48,731538	9,150594

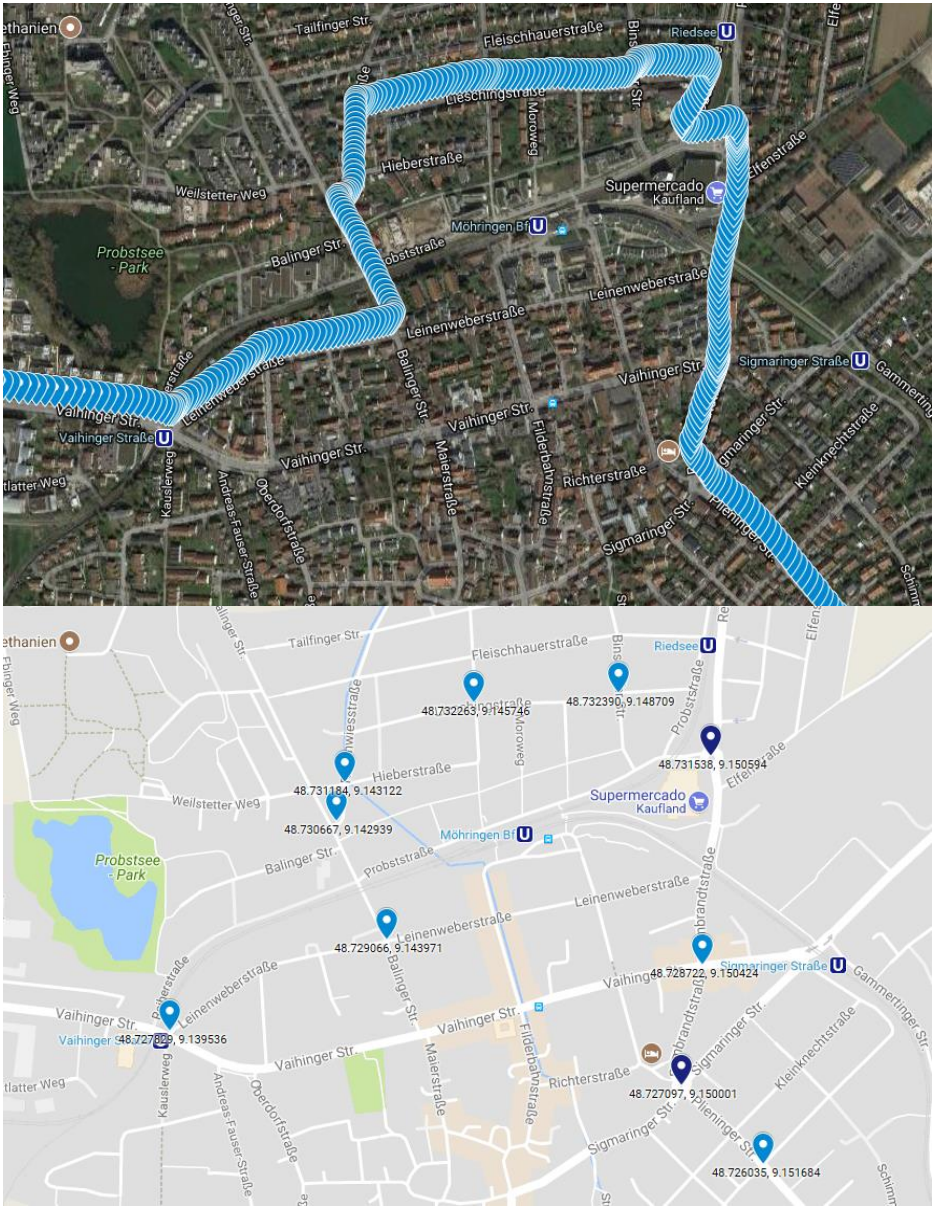

Figura 4.8 Cruces y rotondas en el tramo urbano de Stuttgart

En la Tabla 4.22 se detallan el número de atípicos totales por cada elemento urbano del total de las 10 conducciones, junto con la velocidad media de estos atípicos. Para clasificar dichos elementos, las rotondas se han identificado con la clase 0 y los cruces con la clase 1, como se puede ver en la Figura 4.9.

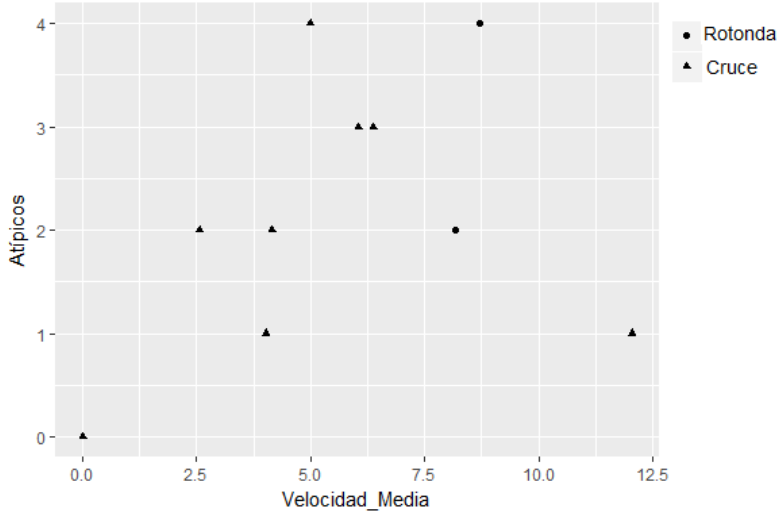

Figura 4.9 Clases de entrada al clasificador en el tramo de Stuttgart

Tabla 4.22 Valores de entrada al clasificador. Prueba 4

Atípicos	Velocidad Media	Clase
4	8,700239	0
2	8,188469	0
0	0,000000	1
1	12,04630	1
2	2,570421	1
3	6,056851	1
3	6,374090	1
2	4,153053	1
4	4,992542	1
1	4,025951	1

Para entrenar los datos de entrada al clasificador, se ha usado de nuevo validación cruzada de *k* iteraciones, tomando un valor de *k* = 5.

La siguiente tabla muestra la matriz de confusión con los resultados obtenidos para la clasificación por ambos clasificadores, donde se da una tasa de acierto del 70%, aunque el coeficiente kappa es prácticamente nulo, y hay que tener en cuenta que las clases no están balanceadas.

Tabla 4.23 Matriz de confusión modelo Logit y SVM. Prueba 4

Real / Predicción	Rotonda	Cruce	Actual	Sensibilidad
Rotonda	0	1	1	0%
Cruce	2	7	9	77,78%
Predicho	2	8	10	77,78% Sensibilidad media
Precisión	0%	87,5%	87,5% Precisión media	70% Exactitud media

4.2. Resultados Experimentales en Escenarios Simulados

Los datos utilizados en estas pruebas, como ya se ha comentado previamente, han sido capturados con un simulador de conducción desarrollado con OpenDS, por el IoTLab de la Facultad de Informática de la Universidad de Reutlingen.

Para la consecución de estos resultados se ha implementado el escenario que se muestra en la Figura 4.10, con un recorrido de unos 4,7 km de longitud, marcado en línea amarilla discontinua, de los cuales 2,2 km corresponden con tramo urbano, mientras que los otros 2,5 km se corresponden a un tramo de carretera. En la figura pueden verse también, indicados mediante círculos, los distintos puntos de interés del recorrido, mostrados en la Tabla 4.24, junto con su distancia aproximada al origen.

Los datos han sido generados por un único conductor y en modo automático, es decir, sin necesidad de meter las marchas, aunque esta opción también está disponible.

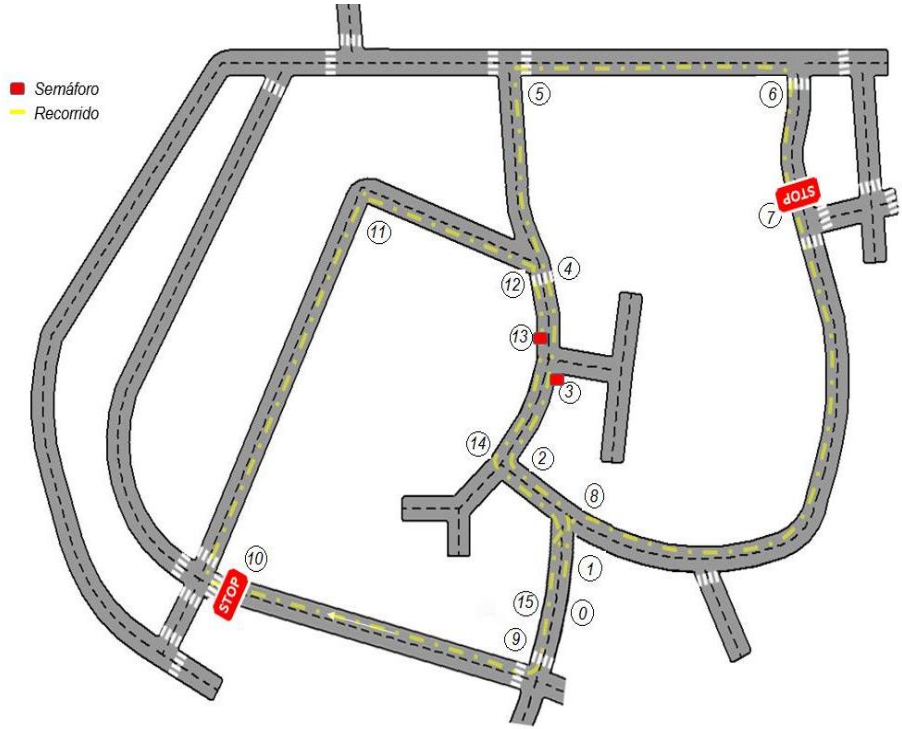

Figura 4.10 Trayecto recorrido en el simulador

4.2.1. Prueba 1S: Cálculo de Atípicos Multivariantes

Nuevamente, en esta primera prueba se va a llevar a cabo una comparación entre las diferentes técnicas de detección de atípicos multivariantes, con los datos capturados bajo simulación por un único conductor y durante 207 recorridos, generando un total de 56.349 observaciones.

Al igual que en las pruebas anteriores, la variable multivariante está formada por las mismas seis variables individuales, cuyos estadísticos correspondientes, así como el número de atípicos univariantes, se muestran en la Tabla 4.25.

En este caso se aprecia que las variables están muy poco correlacionadas, como se puede apreciar en la Tabla 4.26.

Tabla 4.24 Puntos de interés del recorrido

Punto	Descripción	Zona	Distancia al Origen (m)
0	Inicio de trayecto	Urbana	0
1	Primer cruce	Urbana	97,5
2	Segundo cruce	Urbana	180
3	Semáforo 1	Urbana	312
4	Paso de peatones	Urbana	478
5	Tercer cruce	Urbana	588
6	Cuarto cruce	Urbana	839
7	STOP 1	Urbana	1.210
8	Quinto cruce	Urbana	1.458
9	Sexto cruce	Carretera	2.391
10	STOP 2	Carretera	2.686
11	Curva cerrada	Carretera	3.994
12	Séptimo cruce	Urbana	4.223
13	Semáforo 2	Urbana	4.395
14	Octavo cruce	Urbana	4.544
15	Final de trayecto	Urbana	4.695

Tabla 4.25 Estadísticos de las variables univariantes bajo simulación

Variable	Media	Desviación Típica	Mediana	Rango Intercuartílico	Nº de Atípicos
Velocidad Media	17,1856	7,49783	15,1503	11,9777	0
PKE	1,46750	0,68670	1,39196	0,91539	462
Velocidad	17,1852	10,2136	15,0888	13,0222	294
Aceleración	0,00008	2,32722	0,00000	1,96111	4.974
RR	0,82481	0,12819	0,82300	0,16900	880
pNN50	23,9890	13,8479	23,3333	20	796

Tabla 4.26 Matriz de correlaciones. Prueba 1S

	Vel. Media	PKE	Velocidad	Aceleración	RR	pNN50
Vel. Media	1,00000	-0,03771	0,55538	-0,18265	0,10344	0,04259
PKE	-0,03771	1,00000	0,53587	0,21829	-0,02953	0,00974
Velocidad	0,55538	0,53587	1,00000	0,11387	0,06397	0,03653
Aceleración	-0,18265	0,21829	0,11387	1,00000	-0,01581	0,00219
RR	0,10344	-0,02953	0,06397	-0,01581	1,00000	0,22011
pNN50	0,04259	0,00974	0,03653	0,00219	0,22011	1,00000

Los métodos de detección de atípicos multivariantes empleados han sido, una vez más, el algoritmo de Peña y Prieto y MCD, obteniéndose un total de 4.739 y 6.422 atípicos respectivamente.

Además se ha usado el algoritmo de *clustering k-means* para valores de k = 10, 50 y 100, como se muestra en la Tabla 4.27, donde aparecen las 25 observaciones multivariantes atípicas más alejadas, y deja de manifiesto que para estos valores extremos los resultados son muy similares, tanto entre el algoritmo de Peña y Prieto y el método MCD, como para el algoritmo de *clustering*, independientemente del número de *clusters* o grupos.

A continuación, en la Figura 4.11 se puede apreciar una representación univariante de los atípicos multivariantes, es decir, un diagrama de dispersión unidimensional para cada variable, donde se marcan en color rojo los valores atípicos multivariantes.

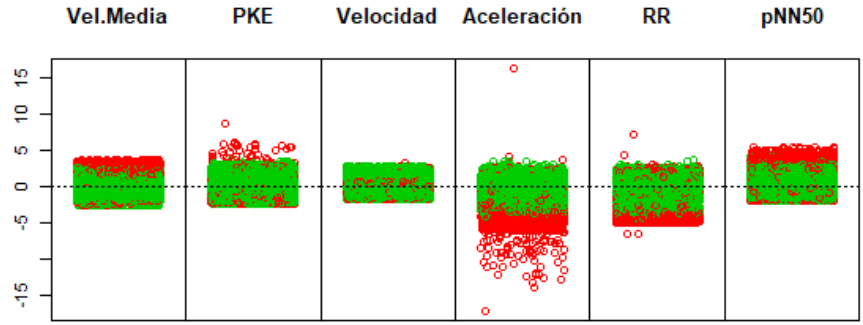

Figura 4.11 Diagrama de dispersión unidimensional de los atípicos multivariantes

Tabla 4.27 Los 25 atípicos multivariantes más alejados obtenidos con los algoritmos de Peña y Prieto, MCD y clustering k-means

Peña y Prieto		MCD	Clustering k-means		
Distancia de Mahalanobis	**Observación**	**Observación**	**k = 10**	**k = 50**	**k = 100**
352,46031	16.819	16.819	6.865	6.865	6.865
299,64628	6.865	6.866	16.819	27.639	27.639
298,60487	6.866	6.865	6.866	6.866	6.866
226,75233	18.690	18.690	27.639	16.819	16.819
205,73070	40.464	40.464	18.690	41.152	41.152
198,51274	6.453	6.453	40.464	27.641	27.641
195,60159	5.533	5.533	16.386	18.690	18.690
192,14330	16.386	26.510	26.510	30.685	30.685
190,99288	26.510	16.386	43.575	27.642	27.642
182,24982	43.575	13.025	5.533	27.645	27.645
176,98375	13.025	40.696	13.025	27.644	40.781
175,29367	40.696	43.575	6.453	40.781	27.644
171,65865	23.748	7.376	40.696	27.643	27.643
170,03830	7.376	23.748	17.971	8.287	9.949
169,57063	27.639	4.386	41.152	2.922	40.780
161,52029	4.386	17.971	23.748	40.780	24.981
157,55649	17.971	27.639	37.263	24.981	43.575
145,69520	6.223	6.223	7.376	9.949	17.971
142,88601	38.425	38.425	27.641	51.671	8.606
140,85190	12.565	12.565	4.386	27.646	43.771
140,23888	37.263	37.263	6.223	18.176	27.646
138,63798	2.950	2.950	30.685	17.903	45.866
135,66014	17.050	17.050	38.425	18.201	17.832
133,80173	9.066	9.066	27.645	45.921	17.075
127,36876	24.133	24.133	17.741	8.606	44.816

4.2.2. Prueba 2S: Detección de Atípicos Multivariantes y Univariantes para Identificación de Puntos de Interés del Recorrido

Al igual que se hizo en las pruebas experimentales con datos reales, la variable univariante a la que se va a aplicar la detección de atípicos es la aceleración del vehículo, calculada como se indicó en el apartado 4.1.3.

Los datos capturados se corresponden con 287 archivos, donde cada archivo contiene las observaciones capturadas por un recorrido completo. En total se han recogido 79.252 observaciones.

En la Tabla 4.28 se muestra el resultado de la detección de atípicos, tanto multivariantes, con el algoritmo de Peña y Prieto, como los atípicos univariantes de la aceleración, donde se indica el número total de atípicos que caen en cada uno de los puntos de interés del trayecto simulado, y cuyas escenas quedan reflejadas en la Figura 4.12.

Tabla 4.28 Número de atípicos identificados para cada punto de interés

Punto	Descripción	Atípicos Univariantes	Atípicos Multivariantes
1	Primer cruce	1	6
2	Segundo cruce	12	3
3	Semáforo 1	237	49
4	Paso de peatones	18	11
5	Tercer cruce	59	9
6	Cuarto cruce	25	4
7	STOP 1	124	5
8	Quinto cruce	93	53
9	Sexto cruce	44	8
10	STOP 2	107	9
11	Curva cerrada	91	32
12	Séptimo cruce	85	12
13	Semáforo 2	96	20
14	Octavo cruce	15	8

Figura 4.12 Escenas de algunos de los puntos de interés de la simulación

En relación a los atípicos univariantes de la aceleración, de la Tabla 4.29 se deduce que los puntos que registran un mayor número de atípicos son los semáforos y las señales de STOP. Esto es así porque en estos puntos el vehículo siempre se detiene. Y los siguientes puntos que presentan más atípicos corresponden a varios cruces o intersecciones, en los cuales se dan diversas situaciones de congestión, además de choques y accidentes, como se puede ver en algunas de las escenas de la Figura 4.12, mientras que otros cruces presentan un número muy inferior de valores atípicos.

En el caso de los atípicos multivariantes los puntos que más atípicos han registrado son de nuevo los cruces donde se suceden la mayoría de incidentes, como atascos y accidentes, seguidos por los semáforos. Sin embargo, esta vez las señales de STOP han generado un número muy pequeño de valores atípicos. Esta situación no es de extrañar y se explica teniendo en cuenta que los más de 200 recorridos han sido realizados por la misma persona, de modo que al conocer el trayecto de memoria solo las situaciones imprevistas generan más apariciones de atípicos, quedando así de manifiesto la influencia de los valores de frecuencia cardíaca.

En cuanto al único paso de cebra que se encuentra en el trayecto, el número de atípicos univariantes y multivariantes es algo similar y bastante pequeño, ya que en este punto, en la mayoría de los recorridos no sucede nada. De vez en cuando es atravesado por un peatón, aunque acontece solo en aproximadamente un 10% de los recorridos.

En el Anexo se recopila un registro del número total de atípicos univariantes encontrados para cada uno de los 287 archivos. Además, se indican también los atípicos multivariantes para cada uno de los archivos que contienen valores de frecuencia cardíaca, que son en total 207 archivos.

Para identificar los atípicos correspondientes a cada uno de los puntos de interés, se toma la distancia de cada uno de ellos con respecto al punto inicial del recorrido. De nuevo se han utilizado los clasificadores Logit y SVM con kernel lineal.

En la Tabla 4.29 se muestra el archivo de datos que alimenta a los clasificadores. Las características extraídas han sido el número de atípicos que caen en cada punto, y la velocidad media y aceleración media que presentan dichos atípicos. La clase cero se corresponde con las intersecciones de 90º, y la clase uno con los cruces. Y se ha tomado un valor de $k = 5$, para la implementación de validación cruzada de k iteraciones.

La matriz de confusión para ambos clasificadores aparece en la Tabla 4.30, con una tasa de acierto del 70%, y un coeficiente kappa de 0,4.

Tabla 4.29 Valores de entrada al clasificador. Prueba 2S

Atípicos	Velocidad Media	Aceleración Media	Clase
59	8,233427	-5,207486	0
25	10,85011	-5,226	0
44	16,30461	-6,744949	0
107	9,763681	-7,688474	0
91	14,23111	-5,505678	0
1	13,12222	-0,8555556	1
12	6,8875	-2,837269	1
93	8,379301	-5,399194	1
85	10,05765	-5,457614	1
15	6,488333	-3,122963	1

Tabla 4.30 Matriz de confusión modelo Logit y SVM. Prueba 2S

Real / Predicción	Intersección	Cruce	Actual	Sensibilidad
Intersección	3	1	4	75%
Cruce	2	4	6	66,66%
Predicho	5	5	10	70,83% Sensibilidad media
Precisión	60%	80%	70% Precisión Media	70% Exactitud media

5. DISCUSIÓN Y CONCLUSIONES

LA labor realizada en este trabajo presenta un novedoso mecanismo para la detección automática de situaciones anómalas de tráfico y la identificación de diferentes elementos de la infraestructura vial. La investigación se basa en la aplicación de técnicas estadísticas y de aprendizaje automático sobre datos de sensores recogidos durante la conducción.

El modelo propuesto combina el uso de métodos de detección de atípicos y clasificadores, sobre datos generados a partir de las ubicaciones estimadas por GPS y velocidades instantáneas, además de una banda de frecuencia cardíaca.

5.1. Discusión de los Resultados

En este apartado se comentan y discuten los resultados obtenidos. En primer lugar se comentarán los resultados de todas las pruebas experimentales realizadas con datos de tráfico real, y a continuación se comentarán los resultados llevados a cabo bajo un entorno de simulación.

5.1.1. Pruebas Experimentales en Escenarios Reales

En la primera prueba se han evaluado y comparado cinco algoritmos de detección de atípicos multivariantes, sobre valores generados a partir de los patrones de velocidad y aceleración, así como los patrones de variación de frecuencia cardíaca.

Por un lado, los resultados obtenidos ponen de manifiesto que los métodos paramétricos o estadísticos utilizados, como son el algoritmo de Peña y Prieto y el algoritmo MCD, los cuales asumen que la variable multivariante se distribuye con una función de probabilidad, normalmente gaussiana, son muy buenos incluso si los datos no se ajustan a ninguna distribución, al menos para las observaciones más extremas.

Por otro lado, queda también reflejado que los algoritmos de minería de datos como pueden ser LOF o *clustering k-means*, obtienen buenos resultados aunque su finalidad no sea explícita la de búsqueda de valores atípicos. Pese a que, en general, los métodos de agrupamiento funcionan peor que los métodos basados en la densidad, en este escenario, donde los datos no se distribuyen como una normal, el algoritmo de agrupamiento o *clustering* parece tener una mejor respuesta.

Y el método más restrictivo corresponde a la SVM de una clase, donde el número total de valores atípicos encontrados es un 10% inferior. Según Petrovskiy (2.003) el algoritmo SVM para detección de atípicos es uno de los más eficientes en minería de datos.

En cuanto al rendimiento de las distintas técnicas, para las estadísticas se tiene un tiempo de cálculo lineal en términos del tamaño de los datos y la dimensión, del orden $O(pn)$, siendo el tiempo de cálculo mayor para el algoritmo de *clustering*, basado en el cálculo de distancias, y mucho mayor en el caso del algoritmo LOF, basado en densidades.

En el caso del algoritmo de Peña y Prieto el tiempo de ejecución para las 3.808 observaciones de esta primera prueba es bastante elevado, teniendo un valor aproximado de 88,19 segundos. No obstante, cabe la posibilidad de utilizar este método en tiempo real, si se calculan los parámetros estimados de la distribución de los datos, esto es, el vector de medias robusto y la matriz de covarianzas robustas. Un vez estimados estos parámetros para los datos capturados por un mismo conductor, es posible inferir si una nueva observación será, o no, atípica, calculando su distancia de Mahalanobis con los parámetros estimados previamente, sin necesidad de ejecutar de nuevo el algoritmo. Pudiéndose generalizar para cualquier conductor si se recoge la suficiente cantidad de información por muchos conductores distintos.

En la segunda prueba se realiza una clasificación supervisada de situaciones de tráfico normal o con retenciones. Para ello se emplean los valores atípicos encontrados en los datos capturados por dos conductores diferentes en un tramo de autovía. Los clasificadores utilizados han sido el modelo Logit y una SVM con kernel lineal.

En primer lugar se ha realizado una clasificación con características extraídas de los valores atípicos y de la velocidad, obteniéndose unos muy buenos resultados, con tasas de acierto del 93,75% y 100%.

En segundo lugar se ha realizado la clasificación con características extraídas únicamente en función de los valores atípicos. Estos atípicos se han calculado con el algoritmo de Peña y Prieto y con una SVM de una clase. Los resultados de la clasificación han sido idénticos para ambos clasificadores, con una tasa de acierto del 96,88%, y muy similares en el segundo con tasas del 93,75% y 96,88%, respectivamente.

Por tanto, en cuanto al rendimiento de los clasificadores, se puede constatar que ambos algoritmos logran resultados muy buenos y bastante similares, aunque el SVM mostró un comportamiento ligeramente mejor.

Sin embargo, estos resultados se han obtenido con un número muy reducido de valores de entrada, por lo que podrían no ser muy significativos, aun usando validación cruzada de *k* iteraciones para aumentar el número de muestras.

En las últimas dos pruebas realizadas con datos de tráfico real se ha trabajado con una única variable, la aceleración del vehículo. A partir de los atípicos obtenidos en el patrón de la aceleración se van detectar ubicaciones candidatas que podrían contener un elemento de la infraestructura vial urbana, como pueden ser rotondas o pasos de cebra.

Nuevamente se han entrenado los clasificadores Logit y SVM con kernel lineal, de forma supervisada, teniendo en cuenta como características el número de atípicos por infraestructura y la velocidad media de estos atípicos.

Una prueba corresponde a un tramo urbano de Leganés donde los datos son generados por un único conductor durante 15 días, con tasas de acierto del 100% y 90% para Logit y SVM respectivamente, mientras que la última prueba se ha realizado con datos recolectados por 10 conductores en un tramo urbano de

Stuttgart. En este caso la tasa de acierto ha sido del 70% para ambos clasificadores, aunque las clases no están balanceadas, ya que se tienen solamente dos rotondas frente a ocho cruces.

Aun cuando los resultados han empeorado teniendo datos generados por diferentes conductores, no implica necesariamente que el rendimiento de los clasificadores sea peor, ya que una vez más la cantidad de datos para entrenar los clasificadores parece insuficiente.

5.1.2. Pruebas Experimentales en Escenarios Simulados

Los datos recolectados bajo simulación han sido capturados por un único conductor durante 287 recorridos, en un trayecto que incluye zonas urbanas y de carretera, generando un total de 56.319 observaciones.

Los algoritmos de detección de atípicos se han aplicado de nuevo a la variable univariante de la aceleración del vehículo, así como a la variable multivariante utilizada en las pruebas con escenarios reales.

En cuanto a la detección multivariante, se han empleado los algoritmos de Peña y Prieto y MCD, donde el número de valores atípicos encontrados por ambos, se muestran en la Tabla 5-1, con una coincidencia del 99,95%.

Para la detección univariante, una vez más, la aceleración, ha sido calculada a partir de los datos capturados de la velocidad. En este caso la velocidad del simulador se encuentra limitada a 180 km/h, que es el límite aplicable en Alemania.

En relación a la prueba realizada para identificación de cruces e intersecciones de 90º, a partir de los atípicos encontrados en la aceleración, el resultado obtenido de la clasificación es una tasa de acierto del 70%, para ambos clasificadores Logit y SVM con kernel lineal.

Tabla 5.1 Número de atípicos multivariantes detectados en entorno de simulación

Técnica	Atípicos/Observaciones	% Atípicos
Algoritmo de Peña y Prieto	4.739/56.319	8,41%
Algoritmo MCD	6.422/56.319	11,40%

Aunque en este caso el coeficiente kappa ha sido 0,4. Se recuerda que este índice es una medida de la diferencia entre la exactitud lograda con un clasificador automático y la probabilidad de lograr una clasificación correcta con un clasificador aleatorio.

En el siguiente anexo se muestran los resultados obtenidos para cada uno de los ficheros generados por cada recorrido. En él, aparece una tabla donde se pueden ver los valores atípicos de la variable aceleración para los 287 recorridos realizados. La tercera columna indica el número de atípicos encontrados en la aceleración y la cuarta columna es el porcentaje total de esos atípicos con respecto al número de observaciones. La quinta columna indica el porcentaje del total de valores atípicos que se ubican en alguno de los 14 puntos de interés señalados en el recorrido. Y las tres últimas columnas indican los mismos parámetros pero para el caso multivariante, donde sólo se ha podido recoger información de frecuencia cardíaca en 207 archivos, debido a varios problemas en la transmisión de estos datos.

Del anexo se puede dilucidar que la cantidad de atípicos encontrados es bastante superior en el caso multivariante, pero el porcentaje de atípicos que se corresponden con las situaciones de interés es mayor en el caso univariante. Para la variable aceleración el número mínimo de atípicos que se dan en un recorrido es de 6, mientras que el máximo es de 46, teniendo un 2,4% y 16,18% de porcentaje de atípicos en relación al número de observaciones de cada fichero. Y el porcentaje de esos atípicos que recaen en los puntos de interés varía desde un 0% hasta un máximo de 57,14%. Mientras que para la detección multivariante el número de atípicos varía desde 1 hasta un máximo de 132, con porcentajes que van desde el 0,38% hasta el 38,90%, con respecto al número de observaciones. En cuanto al porcentaje de atípicos que se ubican en los diferentes puntos de interés el porcentaje oscila entre el 0% y 35,7%.

Por tanto parece factible afirmar que mediante la detección de observaciones atípicas en datos medidos durante la conducción por sensores, se pueden localizar elementos de interés, y además de una forma muy simple.

Por último señalar, que en todos los análisis y test llevados a cabo se han obviado toda clase de errores, provenientes tanto desde la etapa de adquisición de datos, como del análisis y tratamiento, como pueden ser errores de transmisión o de sincronización, errores por redondeo de cálculos o pérdida de información por falta

de memoria, entre otros, sin que con ello se vea significativamente afectado el resultado del modelo propuesto.

5.2. Conclusiones

En este trabajo se ha abordado el desafío que supone la búsqueda de valores atípicos en bases de datos multidimensionales, evaluando diferentes técnicas, y usándolas en combinación con métodos de clasificación supervisada, para la identificación de situaciones anómalas de tráfico y elementos de la infraestructura vial.

En primer lugar, respecto al análisis multivariante, resaltar que cuanto mayor sea la dimensión del espacio de parámetros se necesitará un mayor número de observaciones para que las estimaciones sean fiables. Por tanto para poder detectar, de forma fiable, valores atípicos en una variable multivariante es necesario tener un número de observaciones muy elevado.

Además es necesario considerar la forma y estructura de los datos en el espacio multidimensional, así como todas las dependencias entre variables. Y aunque en todas las pruebas se ha usado el mismo umbral para seleccionar los valores atípicos, no hay ninguna razón de peso por la que este umbral sea fijo y debería ser apropiadamente ajustado para cada conjunto de datos.

En segundo lugar, en cuanto a las pruebas realizadas, todas presentan buenos resultados, aunque todo parece indicar que el número de datos con los que se ha trabajado es insuficiente, así como el número de conductores. Por lo que sería conveniente recopilar mucha más información de distintos conductores, con diferentes trayectos y más elementos.

No obstante, con este trabajo se ha comprobado que el uso de técnicas de detección de atípicos en datos viarios, en combinación con técnicas de clasificación, permiten identificar situaciones anómalas de tráfico, así como de elementos de señalización e infraestructura vial.

Y con todos los resultados obtenidos se puede concluir que se trata de un enfoque bastante prometedor y novedoso en el uso de la detección de atípicos, no tratado antes en esta área.

5.3. Trabajos Futuros

A la vista de las conclusiones obtenidas queda abierta la puerta a numerosas tareas prospectivas, que no se han podido abordar en esta tesis, principalmente, por falta de tiempo.

Antes que nada, como ya ha sido mencionado, sería necesario recopilar una cantidad masiva de datos, en diferentes escenarios y con el mayor número posible de conductores diferentes, de manera que los algoritmos empleados puedan ser validados y mejorados. Además, sería conveniente probar con otros algoritmos distintos, tanto para detección de atípicos, como para realizar las tareas de clasificación.

Los trabajos futuros deberían centrarse en encontrar un conjunto de variables con mayor poder de discriminación, e intentar conseguir información de otro tipo de sensores, como puede ser el acelerómetro, disponible en la mayoría de los teléfonos móviles, junto con la información del GPS. Así, gracias a una información más variada se podrá mejorar la precisión en los métodos de detección de valores atípicos.

Otro enfoque que se considerará en futuros trabajos será incluir otros escenarios reales y/o de simulación para mejorar el sistema y la implementación del modelo propuesto, con la idea de desarrollar algoritmos para predecir situaciones futuras basadas en los datos actuales recibidos por los sensores. La predicción de condiciones futuras del estado de la carretera puede ayudar a elegir la mejor ruta antes de encontrar un posible incidente de tráfico.

Y una de las propuestas más desafiantes podría ser profundizar en el desarrollo de modelos de predicción que puedan extrapolar a distintos escenarios.

Otra posible acción futura sería el diseño e implementación de un sistema que pueda ser usado para soportar otras aplicaciones o servicios de la *SmartCity*.

En cuanto a la etapa de clasificación, sería interesante probar con diferentes vectores de características para entrenar los clasificadores, de manera que se consiga una mayor capacidad de discriminación y comprobar así si se produce alguna mejora en la tasa de acierto. Cabe reseñar que no existe una regla fija para seleccionar las mejores características, luego esto es un proceso arduo que requiere de tiempo, observación y experimentación.

Además sería interesante poder contar con el suficiente número de datos para implementar un aprendizaje no supervisado, ya que no siempre se dispone, a priori, de datos clasificados por clases. Y se podría realizar una evaluación entre los métodos de clasificación supervisada frente a la clasificación no supervisada.

Otra línea de trabajo futuro sería integrar en el diseño distintas herramientas que podrían ayudar a llevar a cabo análisis estadísticos más avanzados.

6. PUBLICACIONES

EN este último capítulo se muestran las publicaciones que han sido generadas, hasta el momento, durante el proceso de elaboración de este proyecto de investigación, y donde se han divulgado las propuestas teóricas, dentro del marco de trabajo planteado.

A continuación, en el siguiente apartado se muestra una lista de los trabajos que han sido publicados en revistas científicas indexadas en el JCR *(Journal Citation Reports)*, sobre la base de los hallazgos encontrados en esta tesis. Seguido de un breve resumen de los resultados publicados.

6.1. Listado de Publicaciones

- *'Evaluation of Outliers Detection Algorithms for Traffic Congestion Assessment in Smart City Traffic Data from Vehicle Sensors'*

  Ramona Ruiz-Blázquez, Mario Muñoz-Organero, Luis Sánchez-Fernández. International Journal of Heavy Vehicle Systems. Inderscience Publishers Ltd., 2018

  Categoría: JCR Q4 (2016)
  Factor de impacto: 0.308
  Posición relativa: 33/34
  Disciplina: *Transportation, Science & Technology*

- *'Automatic detection of traffic lights, street crossings and urban roundabouts combining outlier detection and deep learning classification techniques based on GPS traces while driving'*

  Mario Muñoz-Organero, Ramona Ruiz-Blázquez, Luis Sánchez-Fernández. Computers, Environment and Urban Systems, Elsevier, vol. 68, p. 1- 8, 2018

  Categoría: JCR Q1 (2016)
  Factor de impacto: 2.659
  Posición relativa: 23/105
  Disciplina: *Environmental Studies*

- *'Detecting different road infrastructural elements based on the stochastic characterization of speed patterns'*

  Mario Muñoz Organero & Ramona Ruiz Blázquez. Journal of Advanced Transportation – July 2017

  Categoría: JCR Q3 (2016)
  Factor de impacto: 1.813
  Posición relativa: 18/34
  Disciplina: *Transportation, Science & Technology*

6.2. Resultados de la Tesis en Publicaciones

En el primer artículo presentado, se han publicado los resultados de las pruebas 1 y 2 que han sido descritas en el capítulo 4, en los apartados 4.1.1. y 4.1.2.

En cuanto al primer test, relacionado con el cálculo de atípicos multivariantes, los resultados del artículo difieren respecto a los de la tesis ya que el número de datos recogidos en la fecha de envío del artículo era inferior. Para el *paper* se obtuvieron datos de conducciones durante 9 días y en la tesis se ha aumentado hasta 15 días. No obstante, ambos resultados son equiparables, donde se tiene que el mayor número de atípicos multivariantes se obtiene con el algoritmo de Peña y Prieto, y el menor con la SVM de una clase, y siguen manteniéndose algunas de las observaciones atípicas más alejadas.

Para el segundo test, de clasificación de atascos, los resultados son los mismos, correspondientes a datos recogidos durante 32 días. En esta prueba no se han utilizado las variables procedentes del sensor de frecuencia cardíaca. Y se han obtenido resultados muy prometedores en relación a la clasificación supervisada en función de características extraídas de las observaciones atípicas.

En el segundo artículo publicado se aborda la identificación de elementos de señalización en vías de circulación urbanas, como son, semáforos, cruces y rotondas, mediante el mecanismo de detección automática formado por un detector de atípicos en combinación con técnicas de clasificación. En este caso las variables empleadas para la obtención de atípicos han sido la velocidad y la aceleración. La detección de atípicos se ha efectuado usando la distancia de Mahalanobis y la clasificación no supervisada se ha llevado a cabo con técnicas de '*Deep Learning'*,para la fase de entrenamiento y clasificadores *k*-NN y SVM, obteniéndose unos resultados promedio del 89% para el *recall* y 88% de precisión.

Se han utilizado dos conjuntos diferentes de datos para validar el algoritmo. El primero es capturado por la aplicación *SmartDriver*, por un único conductor, durante 55 recorridos iguales de 8,1 km, que incluyen dos áreas urbanas y una autopista. El segundo conjunto de datos corresponde al usado en la prueba 4.1.4 (Schneegass et al.,2013).

En el tercer artículo también se presenta un método para la identificación automática de los elementos de señalización correspondientes con semáforos, cruces y rotondas, basada en funciones de probabilidad de masas, medida en cada ubicación. Cada elemento va a quedar caracterizado por un vector de velocidades medidas cuando un conductor lo atraviesa.

En esta propuesta no se va a realizar detección de atípicos, si no que se desarrolla un nuevo algoritmo que utiliza la variación total de la distancia. Ésta se usa para encontrar la similitud entre diferentes puntos de interés y proporciona una medida acerca de cómo dos lugares son estocásticamente similares, basado en patrones estocásticos de la velocidad.

Los datos han sido recogidos por la aplicación *SmartDriver* durante 55 recorridos, en dos tramos urbanos, uno en la ciudad de Leganés, de unos 2,9 km, y el otro en la

ciudad de Getafe, con 1,2 km.

Y para realizar la clasificación no supervisada se emplea un enfoque basado en *k* - NN, obteniéndose unos resultados del 75% de precisión.

ANEXO

Nombre	Nº de Observ.	Nº de Atípicos	% Atípicos	% Ptos. Interés	Atípicos Multiv.	% Atípicos	% Ptos. Interés
2017-12-08_10-43-55	337	22	6,52%	0%	128	37,98%	3,9%
2017-12-08_11-46-45	347	30	8,64%	0%	41	11,81%	4,87%
2017-12-08_12-11-02	442	42	9,5%	4,76%	—	—	—
2017-12-08_12-50-11	418	38	9,09%	0%	—	—	—
2017-12-08_13-05-53	368	27	7,33%	0%	—	—	—
2017-12-08_15-56-16	302	29	9,60%	3,44%	24	7,94%	4,16%
2017-12-08_16-10-53	324	27	8,33%	7,40%	7	2,16%	28,57%
2017-12-08_16-25-26	314	35	11,14%	11,42%	10	3,18%	10%
2017-12-11_10-04-41	294	15	5,1%	6,66%	38	12,92%	7,89%
2017-12-11_10-18-01	269	23	8,55%	30,43%	44	16,35%	6,81%
2017-12-11_10-38-31	302	20	6,62%	30%	41	13,57%	9,75%
2017-12-11_10-45-35	287	38	13,24%	26,31%	8	2,78%	0%
2017-12-11_10-52-04	301	20	6,64%	35%	9	2,99%	0%
2017-12-11_10-58-37	323	33	10,21%	12,12%	118	36,53%	5,93%
2017-12-11_11-06-05	316	33	10,44%	24,24%	10	3,16%	0%
2017-12-11_11-35-16	302	29	9,6%	27,58%	—	—	—
2017-12-11_11-52-49	301	26	8,63%	26,92%	—	—	—
2017-12-11_11-59-16	322	28	8,69%	10,71%	—	—	—
2017-12-12_10-26-35	296	21	7,09%	23,8%	—	—	—

Nombre	Nº de Observ.	Nº de Atípicos	% Atípicos	% Ptos. Interés	Atípicos Multiv.	% Atípicos	% Ptos. Interés
2017-12-12_10-46-31	306	28	9,15%	21,42%	—	—	—
2017-12-12_10-57-06	296	28	9,46%	28,57%	—	—	—
2017-12-12_11-16-39	285	27	9,47%	11,11%	—	—	—
2017-12-12_11-22-40	281	28	9,96%	28,57%	—	—	—
2017-12-12_11-29-00	291	43	14,77%	6,97%	—	—	—
2017-12-12_11-35-47	291	21	7,21%	42,85%	—	—	—
2017-12-12_11-42-04	272	17	6,25%	11,76%	—	—	—
2017-12-12_12-03-57	291	34	11,68%	20,58%	23	7,9%	4,34%
2017-12-12_12-10-22	283	22	7,77%	22,72%	16	5,65%	0%
2017-12-12_14-27-08	281	18	6,4%	0%	7	2,49%	14,28%
2017-12-12_14-34-12	293	28	9,55%	7,14%	43	14,67%	2,32%
2017-12-12_14-42-23	247	6	2,43%	0%	14	5,66%	0%
2017-12-12_14-47-50	237	13	5,48%	23,07%	9	3,79%	11,11%
2017-12-12_14-55-00	272	22	8,08%	18,18%	11	4,04%	0%
2017-12-12_15-01-03	286	19	6,64%	26,31%	28	9,79%	3,57%
2017-12-12_15-07-30	280	21	7,5%	28,57%	101	36,07%	6,93%
2017-12-12_15-13-55	275	19	6,9%	42,1%	101	36,72%	13,86%
2017-12-12_15-20-00	284	25	8,8%	8%	23	8,09%	0%
2017-12-12_15-27-14	286	23	8,04%	17,39%	8	2,79%	0%
2017-12-12_15-39-56	288	19	6,59%	5,26%	31	10,76%	9,67%
2017-12-12_16-26-00	285	17	5,96%	47,05%	—	—	—
2017-12-12_16-38-53	264	17	6,44%	41,17%	3	1,13%	0%
2017-12-12_16-47-04	272	16	5,88%	12,5%	—	—	—
2017-12-12_16-53-21	264	20	7,57%	5%	—	—	—
2017-12-12_16-59-13	298	22	7,38%	27,27%	—	—	—
2017-12-12_17-06-14	285	21	7,36%	4,76%	—	—	—
2017-12-12_17-12-34	279	14	5,01%	35,71%	—	—	—
2017-12-13_09-55-45	298	32	10,73%	6,25%	29	9,73%	0%
2017-12-13_10-04-25	294	21	7,14%	4,76%	97	32,99%	5,15%
2017-12-13_10-10-49	293	31	10,58%	16,13%	15	5,12%	0%

Nombre	Nº de Observ.	Nº de Atípicos	% Atípicos	% Ptos. Interés	Atípicos Multiv.	% Atípicos	% Ptos. Interés
2017-12-13_10-17-07	251	11	4,38%	9,09%	2	0,79%	0%
2017-12-13_10-34-17	277	19	6,86%	5,26%	103	37,18%	3,88%
2017-12-13_10-40-20	304	22	7,23%	0%	116	38,15%	4,31%
2017-12-13_10-47-44	298	25	8,39%	0%	12	4,02%	0%
2017-12-13_11-13-56	261	23	8,81%	30,43%	—	—	—
2017-12-13_11-20-10	302	26	8,6%	30,77%	—	—	—
2017-12-13_11-26-44	267	25	9,36%	28%	—	—	—
2017-12-13_11-32-28	279	17	6,09%	29,41%	—	—	—
2017-12-13_11-39-00	282	29	10,28%	31,03%	—	—	—
2017-12-13_11-45-07	243	22	9,05%	22,72%	—	—	—
2017-12-13_12-08-17	272	25	9,19%	24%	—	—	—
2017-12-13_12-14-27	275	16	5,81%	12,5%	—	—	—
2017-12-13_12-20-34	261	13	4,98%	0%	—	—	—
2017-12-13_12-26-35	283	36	12,72%	13,88%	—	—	—
2017-12-13_12-33-21	284	14	4,93%	57,14%	—	—	—
2017-12-13_12-39-23	276	27	9,78%	22,22%	—	—	—
2017-12-13_13-00-40	262	18	6,87%	5,55%	—	—	—
2017-12-13_14-17-02	272	26	9,55%	3,84%	—	—	—
2017-12-13_14-24-34	270	26	9,63%	0%	—	—	—
2017-12-13_14-31-03	276	20	7,24%	0%	—	—	—
2017-12-14_10-17-38	307	37	12,05%	0%	52	16,93%	3,84%
2017-12-14_10-29-51	283	17	6%	5,88%	56	19,78%	1,78%
2017-12-14_10-35-50	241	19	7,88%	10,52%	31	12,86%	3,22%
2017-12-14_10-41-42	284	22	7,74%	9,09%	47	16,54%	2,12%
2017-12-14_10-53-25	255	21	8,23%	23,8%	66	25,88%	7,57%
2017-12-14_10-59-06	250	20	8%	30%	38	15,2%	5,26%
2017-12-14_11-05-14	262	23	8,77%	26,08%	27	10,3%	0%
2017-12-14_11-10-57	249	20	8,03%	25%	29	11,64%	17,24%
2017-12-14_11-22-31	255	19	7,45%	26,31%	27	10,58%	7,4%
2017-12-14_12-15-00	292	33	11,3%	9,09%	—	—	—

Nombre	Nº de Observ.	Nº de Atípicos	% Atípicos	% Ptos. Interés	Atípicos Multiv.	% Atípicos	% Ptos. Interés
2017-12-14_12-21-20	271	21	7,75%	4,76%	—	—	—
2017-12-14_12-27-40	266	22	8,27%	13,63%	2	0,75%	0%
2017-12-14_12-34-07	246	15	6,09%	6,66%	—	—	—
2017-12-14_15-43-11	245	22	8,98%	9,09%	—	—	—
2017-12-14_15-49-17	249	17	6,82%	17,64%	—	—	—
2017-12-14_16-00-14	242	18	7,43%	27,77%	—	—	—
2017-12-14_16-05-31	242	16	6,61%	18,75%	—	—	—
2017-12-14_16-11-16	245	21	8,57%	0%	—	—	—
2017-12-14_16-16-47	244	22	9,01%	22,72%	—	—	—
2017-12-14_16-22-44	251	17	6,77%	29,41%	—	—	—
2017-12-14_16-28-19	233	14	6%	14,28%	—	—	—
2017-12-14_16-34-42	256	25	9,76%	4%	—	—	—
2017-12-14_16-40-24	270	31	11,48%	0%	—	—	—
2017-12-14_16-46-27	248	19	7,66%	15,79%	—	—	—
2017-12-14_17-33-46	238	29	12,18%	13,79%	—	—	—
2017-12-15_10-16-41	249	18	7,22%	16,66%	—	—	—
2017-12-15_10-31-16	254	24	9,44%	37,5%	—	—	—
2017-12-15_10-37-00	247	19	7,69%	26,31%	33	13,36%	9,09%
2017-12-15_10-49-17	259	17	6,56%	11,76%	47	20,43%	4,25%
2017-12-15_11-00-53	256	25	9,76%	36%	22	9,48%	18,18%
2017-12-15_11-06-31	262	18	6,87%	11,11%	—	—	—
2017-12-15_11-25-40	262	24	9,16%	16,66%	9	3,89%	0%
2017-12-15_11-32-35	259	23	8,88%	17,39%	18	7,82%	11,11%
2017-12-15_11-38-25	297	33	11,11%	18,18%	33	11,11%	6,06%
2017-12-15_11-44-53	324	28	8,64%	21,42%	121	38,9%	7,43%
2017-12-15_11-51-39	334	28	8,38%	17,85%	102	30,53%	4,9%
2017-12-15_11-58-35	349	31	8,88%	32,25%	45	12,89%	8,88%
2017-12-15_12-05-55	358	35	9,77%	5,71%	49	13,68%	6,12%
2017-12-15_12-14-09	307	39	12,7%	7,69%	14	4,56%	0%
2017-12-15_12-20-40	246	19	7,72%	26,31%	27	11,63%	0%

Nombre	Nº de Observ.	Nº de Atípicos	% Atípicos	% Ptos. Interés	Atípicos Multiv.	% Atípicos	% Ptos. Interés
2017-12-15_12-26-28	278	33	11,87%	9,09%	25	10,72%	0%
2017-12-15_12-33-00	267	27	10,11%	22,22%	6	2,57%	0%
2017-12-15_12-38-51	272	20	7,35%	30%	8	2,94%	25%
2017-12-15_12-47-23	319	25	7,83%	12%	9	2,89%	0%
2017-12-15_12-54-34	292	21	7,19%	0%	74	25,34%	5,4%
2017-12-15_13-01-41	266	28	10,52%	17,85%	26	11,15%	3,84%
2017-12-15_13-07-34	250	16	6,4%	0%	7	3%	0%
2017-12-15_13-15-14	249	18	7,22%	38,88%	15	6,46%	6,66%
2017-12-15_15-02-55	287	24	8,36%	0%	—	—	—
2017-12-15_15-09-46	262	32	12,21%	9,37%	6	2,58%	0%
2017-12-15_15-15-39	256	19	7,42%	5,26%	45	19,39%	0%
2017-12-15_15-21-30	265	17	6,41%	29,41%	23	9,87%	13,04%
2017-12-15_15-27-23	266	15	5,64%	6,66%	9	3,86%	0%
2017-12-15_15-33-29	271	23	8,48%	0%	35	15,08%	0%
2017-12-15_15-39-32	277	35	12,63%	2,85%	10	4,29%	0%
2017-12-15_15-45-32	297	25	8,41%	12%	—	—	—
2017-12-15_15-52-23	281	23	8,18%	4,34%	—	—	—
2017-12-15_15-58-24	300	15	5%	20%	31	10,33%	0%
2017-12-15_16-04-51	286	22	7,69%	18,18%	17	5,94%	5,88%
2017-12-15_16-11-29	281	21	7,47%	19,04%	11	3,91%	0%
2017-12-15_16-26-20	269	24	8,92%	29,16%	7	3,01%	0%
2017-12-15_16-32-35	242	22	9,09%	18,18%	—	—	—
2017-12-15_16-39-57	258	20	7,75%	30%	—	—	—
2017-12-15_16-46-03	255	18	7,05%	11,11%	14	6,03%	35,71%
2017-12-15_16-51-44	272	23	8,45%	34,78%	—	—	—
2017-12-15_16-57-38	271	24	8,85%	16,66%	2	0,73%	0%
2017-12-15_17-03-58	274	23	8,39%	17,39%	—	—	—
2017-12-15_17-10-01	290	24	8,27%	8,33%	—	—	—
2017-12-15_17-16-28	268	18	6,71%	11,11%	—	—	—
2017-12-18_10-03-27	321	28	8,72%	0%	—	—	—

Nombre	Nº de Observ.	Nº de Atípicos	% Atípicos	% Ptos. Interés	Atípicos Multiv.	% Atípicos	% Ptos. Interés
2017-12-18_10-11-28	298	20	6,71%	25%	36	12,08%	11,11%
2017-12-18_10-30-59	295	19	6,44%	21,05%	—	—	—
2017-12-18_10-44-32	297	23	7,74%	30,43%	75	25,25%	6,66%
2017-12-18_10-57-53	283	27	9,54%	18,51%	102	36,04%	1,96%
2017-12-18_11-04-07	268	24	8,95%	25%	32	11,94%	6,25%
2017-12-18_11-10-25	285	18	6,31%	44,44%	24	8,42%	4,16%
2017-12-18_11-24-08	288	22	7,63%	18,18%	30	10,41%	0%
2017-12-18_11-30-34	293	21	7,16%	38,09%	14	4,77%	14,28%
2017-12-18_11-37-00	287	23	8,01%	0%	—	—	—
2017-12-18_11-43-41	287	21	7,31%	14,28%	15	5,22%	0%
2017-12-18_11-50-46	291	20	6,87%	15%	35	12,02%	2,85%
2017-12-18_11-57-56	279	23	8,24%	17,39%	18	6,45%	5,55%
2017-12-18_12-04-24	281	18	6,4%	55,55%	—	—	—
2017-12-18_12-10-35	295	21	7,11%	23,8%	28	9,49%	3,57%
2017-12-18_12-17-07	288	28	9,72%	10,71%	98	34,02%	11,22%
2017-12-18_12-23-21	301	30	9,96%	10%	112	37,2%	9,82%
2017-12-19_09-50-36	371	46	12,39%	2,17%	132	35,57%	3,03%
2017-12-19_10-07-31	297	20	6,73%	10%	105	35,35%	6,66%
2017-12-19_10-15-09	292	32	10,95%	25%	22	7,53%	13,63%
2017-12-19_10-21-33	287	22	7,66%	13,63%	22	7,66%	4,54%
2017-12-19_10-27-53	295	20	6,77%	15%	9	3,05%	0%
2017-12-19_10-37-57	356	32	8,98%	12,5%	101	28,37%	1,98%
2017-12-19_10-45-47	283	27	9,54%	25,92%	14	4,94%	0%
2017-12-19_10-52-30	270	18	6,66%	11,11%	8	2,96%	12,5%
2017-12-19_10-58-39	267	22	8,24%	9,09%	12	4,49%	8,33%
2017-12-19_11-04-55	268	28	10,44%	3,57%	25	10,82%	4%
2017-12-19_11-10-57	279	19	6,81%	15,78%	12	4,3%	0%
2017-12-19_11-17-07	279	28	10,03%	21,42%	12	4,3%	0%
2017-12-19_11-23-25	312	26	8,33%	15,38%	58	18,83%	1,72%
2017-12-19_11-30-42	284	16	5,63%	18,75%	23	8,09%	0%

Nombre	Nº de Observ.	Nº de Atípicos	% Atípicos	% Ptos. Interés	Atípicos Multiv.	% Atípicos	% Ptos. Interés
2017-12-19_11-37-07	251	13	5,18%	15,38%	3	1,3%	0%
2017-12-19_11-43-10	254	13	5,11%	23,07%	6	2,6%	0%
2017-12-19_11-48-57	285	26	9,12%	3,84%	16	5,61%	0%
2017-12-19_11-55-09	276	24	8,69%	20,83%	17	7,39%	0%
2017-12-19_12-01-10	265	19	7,17%	31,57%	10	4,34%	10%
2017-12-19_12-07-13	280	20	7,14%	10%	—	—	—
2017-12-19_14-17-51	304	23	7,56%	4,34%	—	—	—
2017-12-19_14-24-39	261	33	12,64%	18,18%	16	6,92%	0%
2017-12-19_14-30-47	293	36	12,28%	13,88%	33	12,22%	3,03%
2017-12-19_14-37-15	290	33	11,38%	15,15%	93	32,06%	5,37%
2017-12-19_14-43-32	280	26	9,28%	11,53%	16	5,71%	6,25%
2017-12-19_14-49-38	326	32	9,81%	9,37%	112	34,35%	2,67%
2017-12-19_14-56-52	257	21	8,17%	4,76%	—	—	—
2017-12-19_15-02-33	290	23	7,93%	8,69%	12	4,13%	0%
2017-12-19_15-08-55	263	23	8,74%	17,39%	—	—	—
2017-12-19_15-14-46	288	26	9,02%	3,84%	8	2,77%	0%
2017-12-19_15-20-56	327	38	11,62%	10,52%	17	5,19%	0%
2017-12-19_15-28-07	285	29	10,17%	3,44%	13	4,56%	7,69%
2017-12-19_15-34-22	283	20	7,06%	5%	76	26,85%	3,94%
2017-12-19_15-40-30	280	21	7,5%	4,76%	19	7,06%	5,26%
2017-12-19_15-46-38	279	20	7,16%	5%	5	1,79%	0%
2017-12-19_15-52-48	293	18	6,14%	16,66%	24	8,19%	4,16%
2017-12-19_15-59-11	277	20	7,22%	15%	10	3,61%	10%
2017-12-19_16-14-25	287	34	11,84%	8,82%	7	2,43%	0%
2017-12-19_16-20-43	268	15	5,59%	26,66%	2	0,74%	0%
2017-12-19_16-27-06	291	23	7,9%	13,04%	15	5,15%	0%
2017-12-19_16-33-24	273	14	5,12%	21,42%	5	1,83%	0%
2017-12-19_16-52-09	325	37	11,38%	16,21%	27	8,3%	0%
2017-12-19_16-59-57	293	27	9,21%	14,81%	42	14,33%	2,38%
2017-12-19_17-06-26	303	26	8,58%	19,23%	13	4,29%	0%

Nombre	Nº de Observ.	Nº de Atípicos	% Atípicos	% Ptos. Interés	Atípicos Multiv.	% Atípicos	% Ptos. Interés
2017-12-19_17-12-57	294	19	6,46%	10,52%	8	2,72%	0%
2017-12-19_17-19-21	288	20	6,94%	0%	8	2,77%	12,5%
2017-12-20_09-54-45	290	34	11,72%	2,94%	11	3,79%	0%
2017-12-20_10-01-33	280	16	5,71%	12,5%	50	21,73%	8%
2017-12-20_10-07-47	287	15	5,22%	20%	1	0,38%	0%
2017-12-20_10-14-10	268	23	8,58%	4,34%	7	3,04%	0%
2017-12-20_10-20-24	266	16	6,01%	12,5%	2	0,86%	0%
2017-12-20_10-26-30	261	17	6,51%	11,76%	20	8,65%	0%
2017-12-20_10-32-50	232	16	6,89%	12,5%	35	15,21%	2,85%
2017-12-20_10-38-20	262	25	9,54%	24%	40	17,39%	7,5%
2017-12-20_10-44-25	248	17	6,85%	23,52%	3	1,29%	33,33%
2017-12-20_10-50-27	262	19	7,25%	15,78%	6	2,97%	0%
2017-12-20_10-56-40	142	14	9,86%	57,14%	—	—	—
2017-12-20_11-00-46	273	37	13,55%	18,91%	37	16,01%	2,7%
2017-12-20_11-07-06	279	20	7,16%	25%	28	12,12%	0%
2017-12-20_11-13-29	269	17	6,32%	35,29%	34	14,71%	8,82%
2017-12-20_11-19-40	272	14	5,14%	7,14%	18	7,79%	0%
2017-12-20_11-26-08	136	22	16,17%	4,54%	—	—	—
2017-12-20_11-30-37	261	20	7,66%	5%	—	—	—
2017-12-20_11-36-42	300	18	6%	22,22%	16	5,38%	12,5%
2017-12-20_11-43-28	269	24	8,92%	20,83%	36	15,72%	5,55%
2017-12-20_11-56-16	237	10	4,22%	40%	68	30,76%	8,82%
2017-12-21_10-20-45	272	19	6,98%	26,31%	47	17,27%	0%
2017-12-21_10-37-18	244	19	7,78%	10,52%	50	21,64%	2%
2017-12-21_10-48-37	263	16	6,08%	37,5%	14	6,08%	0%
2017-12-21_10-56-24	252	24	9,52%	20,83%	19	8,26%	5,26%
2017-12-21_11-02-20	288	19	6,59%	36,84%	—	—	—
2017-12-21_11-08-42	254	20	7,87%	40%	7	3,04%	0%
2017-12-21_11-14-47	245	13	5,3%	15,38%	2	0,86%	0%
2017-12-21_11-20-23	268	20	7,46%	25%	32	11,94%	0%

Nombre	Nº de Observ.	Nº de Atípicos	% Atípicos	% Ptos. Interés	Atípicos Multiv.	% Atípicos	% Ptos. Interés
2017-12-21_11-26-36	249	22	8,83%	4,54%	21	9,13%	4,76%
2017-12-21_11-32-14	274	24	8,76%	25%	30	13,04%	3,33%
2017-12-21_11-40-26	263	22	8,36%	4,54%	29	12,6%	0%
2017-12-21_11-46-13	261	26	9,96%	26,92%	12	5,21%	8,33%
2017-12-21_11-52-25	269	16	5,94%	18,75%	3	1,11%	0%
2017-12-21_11-58-22	268	19	7.09%	0%	11	4,76%	0%
2017-12-21_13-48-21	293	30	10,23%	0%	18	6,14%	0%
2017-12-21_13-56-42	272	21	7,72%	19,04%	21	9,05%	0%
2017-12-21_14-02-46	277	22	7,94%	9,09%	19	6,85%	0%
2017-12-21_14-08-47	276	19	6,88%	0%	18	7,82%	0%
2017-12-21_14-20-01	274	22	8,03%	40,9%	73	26,64%	6,84%
2017-12-21_14-26-59	278	20	7,19%	30%	58	20,86%	10,34%
2017-12-21_14-33-18	261	17	6,51%	41,17%	22	9,52%	0%
2017-12-21_14-39-06	262	22	8,39%	13,63%	11	4,74%	9,09%
2017-12-21_14-44-56	262	23	8,77%	34,78%	33	12,59%	3,03%
2017-12-21_14-51-52	262	26	9,92%	11,53%	16	6,95%	0%
2017-12-21_15-03-27	253	24	9,48%	0%	36	15,65%	2,77%
2017-12-21_15-09-03	259	16	6,17%	25%	23	10%	0%
2017-12-21_15-15-27	264	21	7,95%	33,33%	51	22,17%	7,84%
2017-12-21_15-21-13	270	18	6,66%	5,55%	15	6,49%	0%
2017-12-21_15-27-00	310	23	7,42%	17,39%	31	10,09%	3,22%
2017-12-21_15-33-39	266	24	9,02%	4,16%	36	15,58%	0%
2017-12-21_15-39-29	265	15	5,66%	53,33%	22	9,52%	0%
2017-12-21_15-45-26	250	11	4,4%	9,09%	16	6,95%	0%
2017-12-21_15-51-05	265	16	6,03%	31,25%	—	—	—
2017-12-21_16-03-31	239	22	9,2%	22,72%	22	9,56%	9,09%
2017-12-21_16-09-51	253	27	10,67%	25,92%	21	9,13%	4,76%
2017-12-21_16-15-51	262	22	8,39%	13,63%	13	5,62%	0%
2017-12-21_16-21-43	258	23	8,91%	8,69%	11	4,74%	0%
2017-12-21_16-33-09	244	16	6,55%	12,5%	13	5,65%	7,69%

Nombre	Nº de Observ.	Nº de Atípicos	% Atípicos	% Ptos. Interés	Atípicos Multiv.	% Atípicos	% Ptos. Interés
2017-12-21_16-38-53	243	15	6,17%	13,33%	34	14,78%	2,94%
2017-12-21_16-44-40	259	24	9,26%	20,83%	23	10%	4,34%
2017-12-21_16-50-30	263	22	8,36%	31,81%	7	3,04%	14,28%
2017-12-21_16-56-29	244	6	2,46%	0%	14	6,06%	0%
2017-12-21_17-01-58	265	22	8,3%	13,63%	13	5,67%	0%
2017-12-21_17-07-52	252	22	8,73%	4,54%	—	—	—
2017-12-21_17-13-26	238	14	5,88%	7,14%	2	0,86%	0%
2017-12-21_17-22-50	227	15	6,6%	33,33%	13	5,72%	7,69%
2017-12-21_17-28-34	248	11	4,43%	9,09%	13	5,62%	0%
2017-12-21_17-34-27	252	18	7,14%	16,66%	15	6,49%	0%
2017-12-21_17-40-08	247	18	7,28%	33,33%	—	—	—
2017-12-22_10-19-44	274	25	9,12%	0%	46	16,78%	4,34%
2017-12-22_10-27-35	235	11	4,68%	27,27%	23	10,08%	17,39%
2017-12-22_10-44-19	242	15	6,19%	40%	44	18,18%	2,27%
2017-12-22_10-50-28	260	16	6,15%	31,25%	27	11,73%	3,7%
2017-12-22_10-56-22	264	22	8,33%	9,09%	25	9,46%	0%
2017-12-22_11-02-11	251	22	8,76%	18,18%	—	—	—
2017-12-22_11-08-01	286	18	6,29%	38,88%	16	5,59%	6,25%
2017-12-22_11-14-19	259	15	5,79%	46,66%	—	—	—
2017-12-22_11-21-29	310	25	8,06%	8%	54	17,58%	3,7%
2017-12-22_11-28-08	271	19	7,01%	15,79%	22	9,56%	4,54%
2017-12-22_11-34-04	270	22	8,14%	9,09%	11	4,78%	0%
2017-12-22_11-40-04	276	16	5,79%	31,25%	14	5,07%	0%
2017-12-22_11-46-13	264	11	4,16%	0%	36	15,65%	0%
2017-12-22_11-52-04	257	12	4,67%	8,33%	11	4,78%	0%
2017-12-22_11-57-43	259	19	7,33%	10,52%	10	4,36%	0%
2017-12-22_12-34-41	235	14	5,95%	7,14%	8	3,4%	0%
2017-12-22_12-45-18	251	26	10,35%	15,38%	20	8,65%	5%

BIBLIOGRAFÍA

Acuña, E. & Rodriguez, C. «A Meta Analysis Study of Outlier Detection Methods in Classification.» *http://academic.uprm.edu/eacuna/paperout*, 2.004

Amer, M., Goldstein, M. and Abdennadher, S. «Enhancing One-class Support Vector Machines for Unsupervised Anomaly Detection.» *Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description,* p. 8-15, Chicago, 2.013

Andrieu, C. & Saint Pierre, G. «Using statistical models to characterize eco-driving style with an aggregated indicator.» *IEEE Intelligent Vehicles Symposium*, p. 63-68, Alcalá de Henares, 2.012

Atkinson, A.C. «Fast Very Robust Methods for the Detection of Multiple Outliers.» *Journal of the American Statistical Association,* vol. 89, nº 428, p. 1329-1339, 1.994

Bacon-Shone, J. & Fung, W.K. «A New Graphical Method for Detecting Single and Multiple Outliers in Univariate and Multivariate Data.» *Applied Statistics,* vol. 36, nº 2, p. 153-162, 1.997

Barnett, V. & Lewis, T. «Outliers in Statistical Data.» *3ed. Chichester, John Wiley & Sons*, 1.994

Barret, B.E. & Gray, J.B. «Leverage, Residual and Interaction Diagnostics for Subsets of cases in Least Squares Regression.» *Computational S Statistics & Data Analysis*, vol. 26, p. 39-52, 1.997

Beckman, R.J. & Cook, R.D. «Outlier … … …s.» *Technometrics*, vol. 25, nº 2, p. 119-158, 1.983

Ben-Gal, I. «Outlier Detection.» *Data Mining and Knowledge Discovery Handbook*, p. 131-146, Springer, 2.005

Breunig, M. M., Kriegel, H.P., Ng, R. T. and Sander, J. «LOF: Identifying Density-Based Local Outliers.» *Proceedings of the ACM SIGMOD International Conference on Management of Data,* p. 93–104, Dallas (TX), 2.000

Chawla, S. and Gionis, A. «K-means-: A Unified Approach to Clustering and Outlier Detection.» *Proceedings of the IAM International Conference on Data Mining* p. 189-197, Texas, 2.013

Chen, S., Wang, W. & Zuylen, H.V. «A Comparison of Outlier Detection Algorithms for ITS Data.» *Expert Systems with Applications*, vol. 37, p. 1169-1178, Elsevier, 2.010

Chourabi, H., Gil-García, J.R., Pardo, T.A., Nam, T., Mellouli, S., Scholl, H.J., Walker, S. & Nahon, K. «Understanding Smart Cities: An Integrative Framework.» *45th Hawaii International Conference on System Sciences,* 2.012

Corcoba Magaña, V. y Muñoz Organero, M. « SmartDriver: An assistant for reducing stress and improve the fuel consumption.» *XVII Jornadas de ARCA*, 2.015

Cortes, C. and Vapnik, V. «Support-Vector Networks.» Machine Learning, vol. 20, p. 1-25, 1.995

Crawley, M. J. «The R Book.» *John Wiley & Sons*, 2.007

Filzmoser, P., Garrett, R.G. & Reimann, C. «Multivariate outlier detection in exploration geochemistry.» Computers & Geosciences, Elsevier, vol. 31, nº 5, p. 579-587, 2.005

Friedman, J. H. and Tukey, J. W. «A Projection Pursuit Algorithm for Exploratory Data Analysis.» *IEEE Transactions on Computers*, vol. 23, nº 9, p. 881-890, 1.974

Ge, Y., Xiong, H., Zhou, Z., Ozdemir, H., Yu, J. and Lee, K. C. «Top-Eye: Top-k Evolving Trajectory Outlier Detection.» *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, p. 1.733–1.736, 2.010

Gnanadesikan, R. & Kettenring, J.R. «Robust Estimates Residuals and Outlier Detection with Multiresponse Data.» *Biometrics*, vol. 28, p. 81-124, 1.972

Gironés Roig, J. «Algoritmos.» *FUOC, Fundación para la Universitat Oberta de Catalunya*, 2.013

Gogoi, P., Bhattacharyya, D.K., Borah, B. & Kalita J.K. «A Survey of Outlier Detection Methods in Network Anomaly Identification.» *The Computer Journal*, vol. 54, nº 4, p. 570-588, 2.011

Grubbs, F.E. «Procedures for Detecting Outlying Observations in Samples.» *Technometrics*, vol. 11, nº 1, p. 1-21, 1.969

Guo, J., Huang, W. and Williams, B.M. «Real time traffic flow outlier detection using short-term traffic conditional variance prediction.» *Transportation Research Part C: Emerging Technologies, Elsevier,* vol. 50, p. 160-172, 2.015

Hair, J.F., Anderson, R.E., Tatham, R.L. & Black, W.C. «Análisis Multivariante.» *5th ed. Madrid Prentice Hall*, 1.999

Hall, R. E., Brookhaven National Laboratory, New York. «The Vision of A

Smart City.» *2nd International Life Extension Technology Workshop*, Paris, 2.000

Hampton, J. R. (2013). «The ECG Made Easy.» *8th ed. Churchill Livingstone, Elsevier, 2.013*

Hastie, T., Tibshirani, R. and Friedman, J. «The Elements of Statistical Learning. Data Mining, Inference, and Prediction.» *2nd ed., Springer*, 2.013

Hawkins, D.M. «Identification of Outliers.» *London, Chapman and Hall*, 1.980

Hawkins, D.M. «The feasible solution algorithm for the minimum covariance determinant estimator in multivariate data.» *Computational Statistics and Data Analysis*, vol. 17, p. 197-210, 1.994

Hawkins, S., He, H., Williams, G., Baxter, R. «Outlier Detection Using Replicator Neural Networks.» *Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery*, p. 170-180, 2.002

Hodge, V.J. & Austin, J. «A Survey of Outlier Detection Methodologies.» *Artificial Intelligence Review*, vol. 22, p. 85-126, Springer, 2.004

Holgado-Barco, A., Riveiro, B., Gonzßlez-Aguilera, D. and Arias, P. «Automatic inventory of road cross-sections frommobile laser scanning system.» *Computer-Aided Civil and Infrastructure Engineering*, vol. 32, no. 1, p. 3–17, 2.017

Hosmer, D.W., Lemeshow, S. «Applied Logistic Regression.» *2nd ed., Wiley*, 2.000

Huber, P.J. «Projection Pursuit.» *The Annals of Statistics*, vol. 13, nº 2, p. 435-475, 1.985

Hubert, M. and Debruyne, M. «Minimum covariance determinant.» *WIREs Comp Stat*, vol. 2, p. 36–43, 2.010

ITU-T Focus Group on Smart Sustainable Cities. «Smart sustainable cities: An analysis of definitions.» 2.014

James, G., Witten, D., Hastie, T. & Tibshirani, R. «An Introduction to Statistical Learning with Applications in R.» *Springer Texts in Statistics*, 2.015

Johnson, R. «Applied Multivariate Statistical Analysis.» *Prentice Hall*, 1.992

Jolliffe, I.T. «Principal Component Analysis.» *2nd ed. Springer*, 2.002

Kahle, D. and Wickham, H. «ggmap: Spatial Visualization with ggplot2.» *The R Journal,* vol. 5/1, 2.013

Karatzoglou, A., Meyer, D. and Hornik, K. « Support Vector Machines in R.» *Journal of Statistical Software*, vol. 19, p. 1-28, 2.006

Knorr, E.M. & Ng, R.T. «Algorithms for Mining Distance-Based Outliers in Large Datasets.» *Proceedings of the 24th VLDB Conference*, p. 392-403, New York, 1.998

Kruskal, J.B. «Toward a practical method which helps uncover the structure of a set of multivariate observations by finding the linear transformation which optimizes a new 'index of condensation'.» *Statistical Computation*, p. 427-440, 1.969

Kuhn, M. «Building predictive models in R using the caret package.» *Journal of Statistical Software,* vol. 28, 2.008

Li, X., Li, Z., Han, J. & Lee, J.G. «Temporal Outlier Detection in Vehicle Traffic Data.» *IEEE International Conference on Data Engineering*, 2.009

Liu, W., Zheng, Y., Chawla, S., Yuan, J. and Xing, X. «Discovering Spatio-temporal Causal Interactions in Traffic Data Streams.» *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data*

Mining, p. 1.010–1.018, 2.011

Ma, M. X., Ngan, H. Y. & Liu, W. «Density-based outlier detection by local outlier factor on largescale traffic data.» *IS&T International Symposium on Electronic Imaging*, 2.016

Manly, B.F.J. «Multivariate Statistical Methods: A Primer.» *2nd ed. London Chapman & Hall*, 1.994

Mardia, K.V., Kent, J.T. and Bibby, J.M. «Multivariate Analysis.» *London: Academic press*, 1.979

Mascetti, S., Ahmetovic, D., Gerino, A., Bernareggi, C., Busso, M. and Rizzi, A. «Robust traffic lights detection on mobile devices for pedestrians with visual impairment.» *Computer Vision and Image Understanding*, vol. 148, p. 123–135, 2.016

Miteus, J.E., Peng, C.K., Henry, I., Goldsmith, R.L. and Goldberger, A.L «The pNNx files: re-examining a widely used heart rate variability measured.» *Heart*, vol. 88, p. 378-380, 2.002

Montanero Fernández, J. «Análisis Multivariante.» *www.unex.es/publicaciones*, 2.008

Münz, G., Li, S. & Carle, G. «Traffic Anomaly Detection Using K-Means Clustering.» 2.007

Muñoz García, J.A. y Amón Uribe, I. «Técnicas para Detección de Outliers Multivariantes.» *Revista en Telecomunicaciones e Informática*, vol. 3, nº 5, p. 11-25, 2.013

Muñoz Organero, M. & Ruiz Blázquez, R. «Detecting different road infrastructural elements based on the stochastic characterization of speed patterns.» *Journal of Advanced Transportation*, 2.017

Muñoz Organero, M. & Ruiz Blázquez, R. «Detecting Steps Walking at very Low Speeds Combining Outlier Detection, Transition Matrices and Autoencoders from Acceleration Patterns.» *Sensors*, vol. 17, nº 10, 2274; 2.017

Muñoz-Organero, M., Ruiz-Blázquez, R. and Sánchez-Fernández, L. «Automatic detection of traffic lights, street crossings and urban roundabouts combining outlier detection and deep learning classification techniques based on GPS traces while driving.» *Computers, Environment and Urban Systems, Elsevier,* vol. 68, p. 1-8, 2.018

Naik, D.N. « Detection of Outliers in the Multivariate Linear Regression Model.» *Commun. Statist. - Theory Meth.*, vol.18, nº 6, p. 2225-2232, 1.989

Penny, K.I. «Apropriate Critical Values when Testing for a Single Multivariate Outlier by Using the Mahalanobis distance.» *Appl. Statist*., vol. 45, nº 1, p. 77-81, 1.996

Penny, K.I. & Jolliffe, I.T. «A Comparison of Multivariate Outlier Detection Methods for Clinical Laboratory Safety Data.» *The Statistician*, vol. 50, nº 3, p. 295-308, 2.001

Peña, D. «Análisis de Datos Multivariantes.» *Mc Graw Hill Interamericana de España, S. A*., 2.002

Peña, D. & Prieto, F.J. «Multivariate Outlier Detection and Robust Covariance Matrix Estimation.» *Technometrics*, vol. 43, nº 3, p. 286-310, 2.001

Petrovskiy, M.I. «Outlier Detection Algorithms in Data Mining Systems.» *Programming and Computer Software*, vol. 29, nº4, p. 228-237, 2.003

Rao, C.R. «The Use and Interpretation of Principal Component Analysis in Applied Research.» *Sankhya*, p. 329-358, 1.964

Rohlf, F.J. «Generalization of the Gap Test for the Detection of Multivariate

Outliers.» *Biometrics*, vol. 31, p. 93-101, 1.975

Rousseeuw, P.J. «Least Median of Squares Regression.» J*ournal of the American Statistical Association*, vol. 79, p. 871-880, 1.984

Rousseeuw, P.J. «Multivariate Estimation with High Breakdown Point.» *Grossmann W, Pflug G, Vincze I, Wertz W, eds. Mathematical Statistics and Applications,* vol. B, Dordrecht: Reidel Publishing Company, p. 283-297, 1.985

Rousseeuw, P.J. & Leroy, A.M. «Robust Regression and Outlier Detection.» *Wiley Series in Probability and Statistics*, 2.003

Rousseeuw, P.J. & Van Driessen, K. «A Fast Algorithm for the Minimum Covariance Determinant Estimator.» *Technometrics*, vol. 41, nº 3, p. 212-223, 1.999

Rousseeuw, P.J. & Van Zomeren, B.C. «Unmasking Multivariate Outliers and Leverage Points.» *Journal of the American Statistical Association*, vol. 85, nº 411, p. 633-639, 1.990

Ruiz-Blázquez, R., Muñoz-Organero, M. and Sánchez-Fernández, L. «Evaluation of Outliers Detection Algorithms for Traffic Congestion Assessment in Smart City Traffic Data from Vehicle Sensors.» *International Journal of Heavy Vehicle Systems. Inderscience Publishers Ltd.* 2018

Schneegass, S., Pfleging, B., Broy, N., Schmidt, A. and Heinrich, F. « A data set of real world driving to assess driver workload.» *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications.* ACM, New York, p. 150-157, 2.013

Schölkopf, B., Williamson, R., Smola, A. and Platt, J. «Support Vector Method for Novelty Detection.» *Advances in Neural Information Processing Systems*, vol. 12, p. 582-588, 1.999

Schwager, S.J. & Margolin, B.H. «Detection of Multivariate Normal Outliers.» *Chichester, John Wiley & Sons*, 1.982

Toppeta, D. «The Smart City Vision: How Innovation and ICT can build smart, "liveable", ustainable Cities.» *The Innovation Knowledge Foundation*, Think! Report 005/2.010

Tukey, J.W. «Exploratory Data Analysis.» Addison-Wesley, 1.977

Van Aelst, S. and Rousseeuw, P. «Minimum volume ellipsoid.» *WIREs Comp Stat*, vol. 1, p. 71–82, 2.009

Watson, H.C., Milkins, E.E., Holyoake, P.A., Khatib, E.T. & Kumar, S. «Modelling Emissions From Cars.» *Proceedings of the 10th Australian Transport Research Forum*, vol. 1&2, p. 87–109, Melbourne, 1.985

Wu, Z., Watanabe, Y. and Ishikawa, M. «Hybrid LED traffic light detection using high-speed camera.» *Proceedings of the 19th IEEE International Conference on Intelligent Transportation Systems (ITSC '16)*, p. 1235–1241, 2.016

Yay, E. «An adaptive and rule based driving system for energy-efficient and safe driving behaviour.» *International Doctoral Dissertation*, Universidad de Sevilla, 2.016

Zhang, J. «Advancements of Outlier Detection: A Survey.» *ICST Transactions on Scalable Information Systems*, vol. 13, p. 1-26, 2.013

Zhu, T., Wang, J. and Lv, W. «Outlier Mining Based Automatic Incident Detection on Urban Arterial Road.» *Proceedings of the 6th International Conference on Mobile Technology, Application & Systems*. Nice, 2.009

WEBS

British Antarctic Survey. https://www.bas.ac.uk/. Última consulta 10/10/2017

Google Maps. https://www.google.com/maps/. Última consulta 30/06/2017

OpenDS. Open source driving simulation. https://www.opends.eu/. Última consulta 22/12/2017

Proyecto HERMES. http://madeirasic.us.es/hermes/. Última consulta 09/10/2016

The R Project for Statistical Computing. https://www.r-project.org/. Última consulta 12/02/2018

'No veo lógico rechazar datos porque parezcan increíbles.'

- Fred Hoyle (1.915-2.001) -

'Los errores causados por los datos inadecuados son mucho menores que los que se deben a la total ausencia de datos.'

-Charles Babbage (1.792-1.871) -

FE DE ERRATAS

Página	Ubicación	Pone	Debería Poner
27	Apartado 2.4.1.4. Primer párrafo, 3ª línea	a diferencia del PCA, no considera que los datos …	a diferencia del PCA, si considera que los datos …
45	Última línea	0,8 y 1 segundo	0,6 y 1 segundo
60	Tabla 4.2, 1ª columna	Observación	Observación
		3.280	3.251
		3.281	3.252
		551	522
		1.257	1.228
		2.562	2.533
		1.459	1.430
		711	682
		3.636	3.607
		2.022	1.993
		3.635	3.606